Genomics of phages with therapeutic potential

Zschach, Henrike

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Genomics of phages with therapeutic potential

Henrike Zschach

30^{th} November, 2017

DTU Bioinformatics
Department of Bio and Health Informatics

Contents

Preface

Preface

This thesis was prepared at the Department of Bio - and Health informatics, at the Technical University of Denmark (DTU) in fulfilment of the requirements for acquiring a Ph.D. degree. It describes the use of genomics to characterize phages in a commerical cocktail as well as sewage samples from different locations around the world, and mathematical modeling to study the factors of phage susceptiblity in *Staphylococcus aureus*. The thesis consists of a general introduction, two research papers and one manuscript in prepartion produced during the period 2014 - 2017.

The work was carried out under the supervision of professor Morten Nielsen as well as the external supervisors Mette Voldby Larsen (CEO of GoSeqIt, formerly associate professor at DTU Systms Biology) and Henrik Hasman (special consultant at Statens Serum Institute, formerly senior researcher at DTU Food). The Ph.D. was funded by DTU.

Lyngby, November 2017
Henrike Zschach

Abstract

Bacteriophages, viruses that prey on bacteria, have been applied since the 1920's to treat and prevent bacterial infection. After the discovery of antibiotics, this route was however largely abandoned. Now, with antimicrobial resistance in human-pathogenic bacteria on the rise and a dire need for alternatives, phage therapy once again takes center stage.

Phage therapy holds the promise of substantial benefits both from the economic as well as the public health perspective but also holds distinct challenges. The aim of this PhD was to address how bioinformatics tools, specifically genomics and mathematical modelling, can be applied to move the field towards a future of actual phage therapy in humans. It is composed of three related research projects.

The first part of this thesis is an introduction to various topics and methods relevant to the research projects that jointedly make up this PhD. Chapters 1 - 3 deal with phages, their use in therapy and the nosocomial pathogen *Staphylococcus aureus*. Following that, Chapter 4 and 5 provide an overview of Next Generation Sequencing as well as commonly employed genomics tools, while Chapter 6 details basics of Machine Learning.

The second part, divided into three chapters, presents the three research projects. In project 1, an important commercial phage cocktail with a long history was sequenced and its component phages analyzed. It was found that the cocktail is composed of at least 23 different phage types, which were present in differing abundances. Some of these phage types were successfully amplified on a collection of in-house bacteria corresponding to the cocktail's stated bacterial targets. Further, no harmful genes were detected in the cocktail.

Project 2 deals with phage communities in sewage by comparing samples from around the world to each other as well as to databases of available phage genomes. It revealed a great diversity in the sequences, many of which were distant from all known phages. The phage content of the different sample locations exhibited a rather stable genomic distance that was not influenced by whether the locations were geographically close or not.

Project 3 had the goal of identifying gene families in the extensive accessory genome of the hospital pathogen *Staphylococcus aureus* that influence its susceptibility to clincal phage preparations. This was done by phage testing a set of patient-derived *S. aureus* isolates against a panel of phage preparations. We then sought to model the results using the bacteria's genetic background as features. Doing so, we built nine models with sufficient explanatory power over the susceptibility outcome and from them identified a set of 167 gene families

relevant for phage susceptibility.

The third part of the thesis consists of conclusive remarks and a critical reflection on how each of these projects has impacted the field and how they are connected as well as pointing out directions for future investigations.

In summary, the work included in this this thesis focuses on applying genomics and mathematical modelling to questions related to phage therapy.

Dansk resumé

Bakteriofager, virus der inficerer bakterier, er blevet anvendt til forebyggelse og behandling af bakterielle infektioner siden 1920'erne. Efter opdagelsen af antibiotika blev denne praksis dog i det store og hele opgivet. Med den kraftige stigning i antibiotikaresistens blandt humane sygdomsfremkaldende bakterier, og det deraf fremkomne akutte behov for alternativer til antibiotika, træder fag-terapi endnu engang frem på hovedscenen.

Fag-terapi bærer potentialet til store økonomiske såvel som sundhedsmæssige fordele, men indeholder også specifikke udfordringer. Formålet med denne PhD var at addressere hvordan bioinformatiske metoder, i særdeleshed genomics og matematisk modellering, kan anvendes til styrkelse af det videnskabelige felt med henblik på en fremtid hvor fag-terapi i mennesker er en realitet. PhD'en er opbygget af tre relaterede forskningsprojekter.

Første del af afhandlingen udgøres af en introduktion til diverse emner og metoder med relevans for de forskningsprojekter, der tilsammen udgør PhD'en. Kapitel 1-3 omhandler fager, deres terapeutiske brug og den nosokomielle patogen *Staphylococcus aureus*. Efterfølgende giver kapitel 4 og 5 et overblik over Next Generation Sequencing samt metoder, der ofte bruges i genomics. Kapitel 6 omhandler basale maskinlæringsprincipper.

Den anden del, opdelt i tre kapitler, præsenterer de tre forskningsprojekter. I projekt 1 blev en vigtig kommerciel fag-cocktail med en lang historik sekventeret, og de enkelte fager, der udgør cocktailen, blev analyseret. Det blev fundet at cocktailen bestod af mindst 23 forskellige fag-typer, som var tilstede i forskellig mængde. Nogle af disse fager blev med succes opformeret v.h.a. en lokal samling af bakterier, der repræsenterede de typer bakterier, som cocktailen var rettet imod. Der blev ikke fundet nogen skadelige gener i cocktailen.

Projekt 2 omhandler fag-samfund i spildevand, hvor prøver fra verden over blev sammenlignet med hinanden og med fag-genomer i databaser. Dette viste en høj diversitet i sekvenserne, hvoraf mange kun lignede de kendte fager meget fjernt. Fag-indholdet i prøverne udgjorde en forholdsvis stabil genomisk forskellighed, der ikke blev påvirket af den geografiske tæthed hvormed prøverne var blevet taget.

Projekt 3 havde til formål at identificere gen-familier i den del af genomet af *Staphylococcus aureus*, der varierer indenfor arten, og som påvirker bakteriens følsomhed overfor kliniske fag-blandinger. Dette blev gjort ved at teste et sæt af *S. aureus* isoleret fra patienter mod et panel af fag-blandinger. Vi forsøgte dernæst at modellere resultaterne i forhold til bakteriernes genetiske baggrund. I denne proces byggede vi ni modeller, der i tilstrækkelig grad kunne

forklare den observerede følsomhed, og fra disse modeller identificerede vi 167 gen-familier med relevans for bakteriernes følsomhed overfor fager.

Den tredje del af denne afhandling udgøres af de afsluttende konklusioner samt en kritisk refleksion over hvilken indflydelse hver af disse projekter har haft på det videnskabelige felt og hvordan de er forbundne. Derudover udpeges retningslinjer for fremtidige undersøgelser.

Summa summarum, det arbejde, der er inkluderet i denne afhandling, fokuserer på anvendelsen af genomics og matematisk modellering til spørgsmål relateret til fag-terapi.

Acknowledgements

The time of my PhD has been a very crowded three years, in a good way, and I am very thankful to a lot of people for making it such a stimulating and rewarding experience.

First of all, I would like to take this chance to thank my primary supervisors Mette Larsen and Morten Nielsen who have both been a great source of guidance and advice. I am very grateful to Mette for continuing to supervise me on her own time for 1.5 years even as she was embarking on her own adventure of starting up a company; as well as to Morten for taking over as the main supervisor on the project halfway through even though it was not within his field of research. Also thanks to Ole Lund who was not officially my supervisor but nonetheless has been very good at giving advice and encouragement and has helped me with supervising student projects.

Furthermore, I would like to extend thanks to my co-supervisor Henrik Hasman and my close collaborators, Henrik Westh from Hvidovre Hospital, Ryszard Międzybrodzki from Hirszfeld Institute, Betty Cutter from Evergreen State College, Marina Goderdzishvili from Eliava Institute and Zemphira Alavidze from Eliava Biopreparations. Thank you for the exciting work we did together and your aid in writing and editing the manuscripts.

The department of Bio- and Health Informatics (formerly Center for Biological Sequence Analysis) has been a good place with many kind colleagues who were always ready to give advice and help out. There are too many people to mention them all here, but suffice to say I have had many nice talks at the coffee machine and it was a great social environment. Especially the local network of PhD students was a great source of support as well as fun. Special thanks to Morten, Mette, Marie and Franzi who helped me edit and proof-read this thesis.

I would like to further thank the members of Ole's and Morten's research groups. Both groups were great scientific working environments and we also had a lot of fun at numerous barbecues, sailing trips and Christmas dinners.

I am also very thankful for the opportunity to travel as much as I did, to get to know so many interesting people and to have worked together with some of them. The phage research field has been exceptionally friendly and welcoming. It is a community I really enjoy being a part of, both on a professional and a personal level.

Not to forget, a big thank you goes out to the administrative staff and local technical support who did a great job at keeping things running as smoothly

as possible while the department was going through some turbulent times.

Thanks also to my family for always supporting me and putting up with me only calling them every 3 to 6 months because I forgot.

In the end, a special thanks to my office mates down in the basement in 061 - the domain of dark humor and negativity. I started this PhD being excited about science and now I ended it being excited about lunch. Thanks guys, I wouldn't have made it without you.

Papers included in the thesis

- **Henrike Zschach**, Katrine G. Joensen, Barbara Lindhard, Ole Lund, Marina Goderdzishvili, Irina Chkonia, Guliko Jgenti, Nino Kvatadze, Zemphira Alavidze, Elizabeth M. Kutter, Henrik Hasman and Mette V. Larsen. *What Can We Learn from a Metagenomic Analysis of a Georgian Bacteriophage Cocktail?* Viruses.

- **Henrike Zschach**, Mette V. Larsen, Henrik Hasman, Henrik Westh, Morten Nielsen, Ryszard Międzybrodzki, Ewa Jończyk-Matysiak, Beata Weber-Dąbrowska and Andrzej Górski. *Host-genomic determinants of phage susceptibility in MRSA*. Submitted to Antibiotics.

- **Henrike Zschach**, Vanessa Jurtz, Barbara Lindhard, Mette V. Larsen, Ksenia Arkhipova, Bas Dutilh, Morten Nielsen, Rene Hendriksen, Frank Aarestrup, Ole Lund. *Phage communities in sewage – A metagenomics cross-country perspective.* Manuscript in preparation.

Papers not included in the thesis

- Julia Villarroel, Kortine Annina Kleinheinz, Vanessa Isabell Jurtz, **Henrike Zschach**, Ole Lund, Morten Nielsen and Mette Voldby Larsen. *HostPhinder: A Phage Host Prediction Tool.* Viruses.

Abbreviations

DNA	Deoxyribo-Nucleic Acid
ANI	Average Nucleotide Identity
NCBI	National Center of Biotechnology Information
CV	Cross Validation
S. aureus	*Staphylococcus aureus*
MRSA	Methicillin-resistant *Staphylococcus aureus*
HI	Hirszfeld Institute of Immunology and Experimental Therapy of the Polish Academy of Science

Part I

Introduction

1 Phages

1.1 Phage biology

Bacteriophages, shortly referred to as phages, are viruses that infect bacteria. They are the most abundant biological entity on the planet, with 10^{31} phage particles estimated in the biosphere [1]. A cartoon of a T4-phage is shown in Figure 1.1.

The two principal lifestyles observed in phages are the lytic and the lysogenic cycle. Both begin with phage adsorption to a suitable host cell and injection of the phage DNA. In the lytic cycle, the host metabolism is taken over by the invading phage DNA and tuned to replicate said DNA as well as transcribe it to the proteins necessary to produce new phage particles. Once the new phages are assembled, the host is lysed. In short, during the lytic cycle, phage progeny is produced and released.

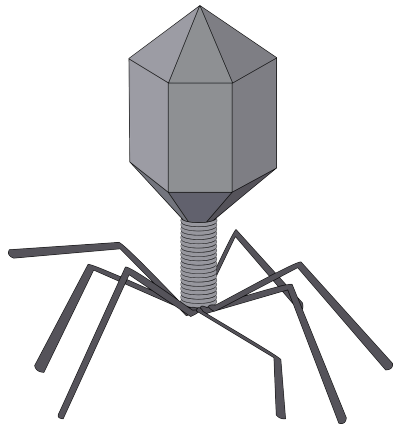


Figure 1.1. Cartoon representation of a T4-phage. It is structurally composed of a capsule or head, a tail shaft and tail fibers.

During the lysogenic cycle however, the phage DNA remains inside the bacterial cell, usually as an integrated prophage or more rarely as a plasmid. It then replicates together with the host cell, effectively creating a new copy of the phage every time the host divides. In this state, the bacterial host may be referred to as a lysogen. An intact prophage may switch back to the lytic cycle

and initiate production of phage progeny and host lysis as described above. It is thought that this switch occurs as a response to stress on the host cell, which can indicate that prospects of survival and further division of the host are unlikely [2].

1.2 Phage taxonomy and genomics

The official authority of phage taxonomy is the International Committee on Taxonomy of Viruses (ICTV). In the broadest context, phages are classified based on their morphology and type of genetic material. Both single-stranded and double-stranded RNA and DNA genomes have been observed, as well as a range of different morphologies, but by far the most common (90%) are tailed phages with double-stranded DNA genomes [3]. Those phages belong to the order of the *Caudovirales*, which can further be subdivided into three families: *Myoviridae*, *Siphoviridae* and *Podoviridae*.

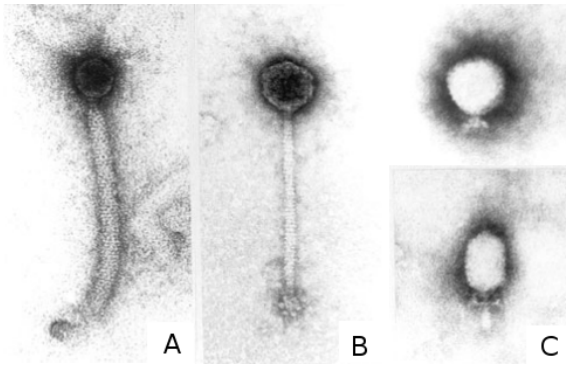


Figure 1.2. Morphology of the three families within the *Caudovirales*. A: *Myoviridae*. B: *Siphoviridae*. C: *Podoviridae*. Adapted from [4].

All *Caudovirales* are tailed phages composed of a capsid or head with cubical symmetry that contains the DNA and a helical tail shaft [3]. Additionally, they often have structures at the end of the tail to facilitate host-recognition and docking, such as base plates, spikes or tail fibers.The three families have distinct tail morphologies and are identified by electron microscopy. *Myoviridae* are marked by long, contractile tails, *Siphoviridae* by long, non-contractile tails and *Podoviridae* by short tails. Examples for each family can be seen in Figure 1.2. This division highlights the major problem of the current classification system: It requires isolation and visualization of the virion. It is

therefore not possible to officially classify phages known only from metagnomic sequencing or prophages identified in bacterial genomes.

As of now, genome-based taxonomy remains difficult because there are no genes shared among all phages that could serve as a marker such as 16s for all cellular life forms [1]. While it is true that bacterial species identification, especially in the epidemiology context, has moved towards sampling a larger proportion of the genome, the sampling of rRNA genes has evolutionized the phylogeny of cellular life by enabling researches to draw a tree of life and place every know lifeform in it. The same is not possible for phages, though efforts have been made to build trees based on overall genome similarity as for example the Phage Proteomic tree by Rohwer and Edwards [1].

Those efforts are hindered by the fact that phage sequences are extremely diverse. This is especially true for phages with non-overlapping host ranges, to the extent where two phages of different hosts seldom share extensive stretches of nucleotides unless they are closely related [5]. In concert with their bacterial hosts, phages have been described to constitute the greatest genetic diversity on earth [6]. The evolution of both phage and bacterial genomes are hugely driven by their interaction with each other, locked in an evolutionary battle of defense-counter-defense mechanism [7].

Phage genomes also range widely in size from 2.4 kb in Leuconostoc phage L5 [5] to $\sim$ 500 kb in Bacillus phage G [6]. Further, their genomes are extensively mosaic, which may be a consequence of frequent horizontal gene transfer [7]. Nonetheless, genes for related functions tend to cluster together into segments [3]. Different segments may have a distinct evolutionary history [7].

In October 2017, 6377 phage genomes were available in NCBI's genbank and 2943 in the phantome database, a dedicated phage resource. There is some overlap across databases. After homology reduction on 100% sequence identity, there are 5570 unique phage genomes known.

In accordance with the enormous diversity described above, the majority of open reading frames found on new phage genomes typically code for proteins with no known function or homolog [8].

2 Phage therapy

The application of phage to treat bacterial infection, commonly referred to as phage therapy, is a promising alternative to antibiotic treatments which are proving increasingly difficult with the spread of antibiotic resistance. Phages can either be used as purified single phage preparations or as cocktails composed of many different phages. Both procedures are in use today [9].

2.1 Bacterial phage resistance mechanisms

As with antibiotics, bacteria may develop resistance towards phage infection. There are several strategies: Evasion of phage recognition, recognition and degradation of phage DNA, general interference with the phage reproductive cycle, and altruistic abortive infection where host cells go into cell death before the phage has finished producing progeny [10].

The first step of phage infection is recognition of and irreversible binding of the phage particle to the host cell. Seeking to evade this recognition by modifying the phage binding site or masking the receptor is an obvious strategy and there are many examples of this in the literature as well as examples of counter-mutation by the phage tail fiber to recognize the altered receptor [11–14].

After successful injection of the phage DNA, the infection can still be stopped by degrading the phage DNA before it takes over the host metabolism. The two most widely-known systems for that are restriction-modification and CRISPR-Cas. Restriction-modification is a 2-component system in which a methylase introduces a specific methylation pattern to the host DNA. DNA that lacks this methylation pattern, i.e. invading foreign DNA, is cut by the accompanying restriction enzyme [10]. Though being a wide-spread phenomenon in bacteria, CRISPR systems are curiously absent in the opportunistic hospital pathogen *Staphylococcus aureus* [15] which is the focal pathogen in this thesis. They will therefore not be described in detail.

Finally, the successful production of phage progeny can be thwarted by the host cell by interfering with one or several steps in the phage replication process. Those systems are referred to as abortive infection systems and, unlike the defense systems described above, result in the death of the host cell [10].

There have also been examples of quorum sensing regulating receptor expression in E. coli and thereby reducing the number of phage infections when it is growing in dense populations [16]. In addition to these bacteria-encoded defense mechanism, acquiring a prophage may protect the bacterial host via the superinfection exclusion system [10].

Despite the plethora of defense mechanisms present, phage therapy can still succeed since, in contrast to antibiotics, phages constantly evolve in concert with their host. Furthermore, there is evidence that the use of cocktails containing complementary phages may reduce the emergence of resistance [17, 18].

2.2 Beginnings

The beginnings of phage therapy go back all the way to the discovery of phages in the late 1910's. In 1915, the Englishman Frederick Twort discovered an agent with bacteriocidal potential on a culture of Staphylococcus. The agent was transferable between cultures and could not be inactivated by Chamberland filtration, meaning it must be extremely small. He published his findings in the Lancet but was unable to follow up on them due to the disruption by First World War. Two years later, in 1917 the Frenchman Felix d'Herelle made similar observations. He went on to perform animal studies as well as human trials to test the potency of this agent, which he dubbed 'bacteriophage' in preventing and mitigating bacterial infection. From there on, therapeutic use of phages quickly expanded during the 1920's [19].

However, controversy about the nature of phages remained and many phage-derived treatments were carried out in poor understanding. Detailed reasons for this are listed by Harper *et al* in a review paper titled 'Phage therapy: Delivering on the promise' [19]. Overall, the supporting evidence for phage treatment was found unconvincing. Phage therapy was therefore deemed inferior to newly discovered antibiotics and was eventually abandoned in the Western world around the 1940's.

2.3 Phage therapy today

Today, phage therapy is almost exclusively available in Russia and Georgia. There are exceptions under the experimental treatment umbrella, see below. In both Russia and Georgia, phage preparations may be purchased as ready-for-use products in pharmacies. The main producers are the companies Microgen (Russia) and Eliava Bio Preparations (Georgia). In this thesis, the focus will

be on Georgian phages.

Eliava Bio Preparation is affiliated with the Eliava Institute, whose roots go back to the very beginning of phage therapy. In 1923, d'Herelle was convinced by his colleague George Eliava to co-found an institute for bacteriophage research in Eliava's native country, Georgia. A photograph of d'Herelle and Eliava working together, presumably taken in Georgia, is shown in Figure 2.1.Though Eliava was later executed and turbulent times followed during the break-up of the Soviet Union, the institute still exists today. It is now known as the George Eliava Institute of Bacteriophages, Microbiology and Virology (Eliava Institute for short) and has accumulated an immense amount of knowledge. The Eliava Institute offers 6 different phage preparations, among them the INTESTI cocktail which has been analyzed in this thesis.

Figure 2.1. Photograph of Felix d'Herelle (mid) and George Eliava (right), ca. 1930's. Taken from the Eliava Institute's website at http://eliavaphagetherapy.com/about-eliava-institute/george-eliava-about-eliava-institute/.

In addition to that, phage therapy is offered to specific cases in the phage therapy unit of the Hirszfeld Institute in Wroclaw, Poland. This use-case is possible as an experimental therapy under the umbrella of the Helsinki declaration, available as a last resort treatment for patients suffering from

chronic, treatment-resistant bacterial infections [20].

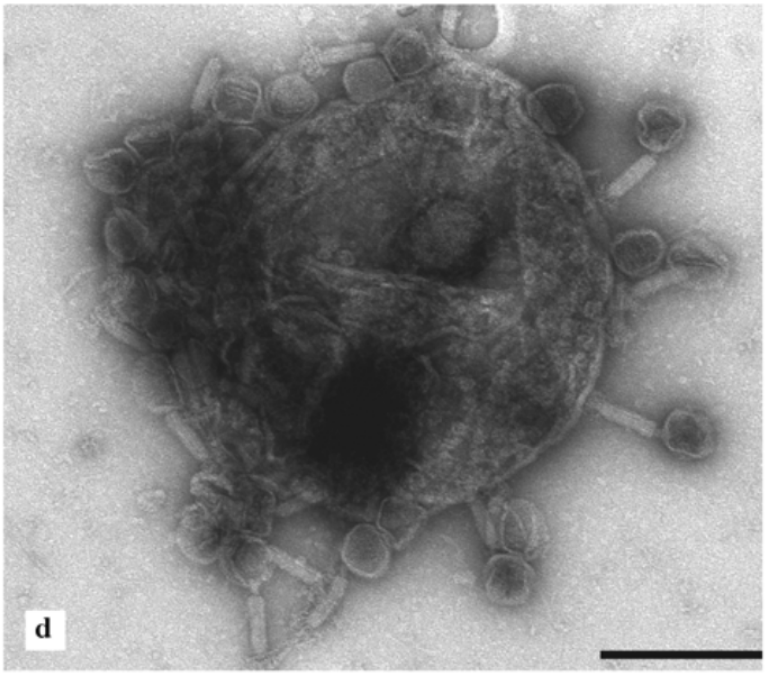


Figure 2.2. Multiple phages of species 'ISP' attached to their host *Staphylococcus aureus*. Bar: 500nm. Reprint from [21].

There are several challenges connected with the use of phage therapy in the Western world. They are both of legislative and regulative as well as practical nature [22]. In the dogma of evidence based medicine a therapeutic should be both effective and safe as well as have a well-characterized mode of action. Phages are generally regarded as a viable solution to the antibiotic crisis by legislation authorities[1]. Their ubiquitous presence in nature and their inherent inability to interact with eukaryotic cells suggests that they should be safe to use in human therapy. However, it holds true that phages carry bacterial virulence factors and in many human-pathogenic bacterial species phages and phage associated mobile genetic elements have been identified as essential to their pathogenicity. It is therefore necessary to thoroughly characterize a candidate therapeutic phage on a genetic level.

[1] As evidenced both by the fact that FDA does give approvals for phages as emergency INDs and by their stated commitment to "facilitating the testing of phage therapy in clinical trials" [23]

Further, for phage therapy to be effective it is necessary to either identify the infecting bacteria down to strain level and test it against a library of phages, or to use a single very broad phage or phage cocktail. In case of the broadband approach the advantage of phages as highly specific agents over antibiotics may be lost. On the practical side, there are questions regarding the mode of delivery since phages are much larger than chemical drugs and it is not clear which sites of the body can be reached effectively by simple oral administration. Another practical consideration is whether phages can induce immune reactions when given in the blood stream as some suggest as mode of delivery.

The way to legislate phage therapy is to go through the legislative channels commonly applied to all medical drugs. However, the very nature of phages as viruses makes them not very suitable for approval criteria that have been designed for chemical drugs, which will not change their composition over time nor be amplified when in contact with their target. Nevertheless, the interest in finding a feasible way to fit phages into the drug legislation is considerable and those challenges will eventually be overcome. There are several initiatives currently underway that aim to provide sufficient evidence regarding efficacy and non-toxicity of phage therapy. Most outstanding is Phagoburn, a phase I-II clinical trial in which a phage preparation is used to treat burn wounds infected with Pseudomonas aeruginosa. It was initiated by the French company Pherecydes Pharma and is being carried out in collaboration with 3 partners and 11 clinical sites (see http://www.phagoburn.eu/). This is a landmark clinical trial that hopefully will aid to pave the way for phage therapy in Europe and the USA.

3 *Staphylococcus aureus*

Staphylococcus is a genus of gram-positive spherical bacteria that grow in grape-like clusters. There are several species, but in humans mainly *S. aureus* and S. epidermidis are clinically relevant [24]. Both species have been found in the normal bacterial flora of healthy individuals with about 20 - 30% of the human population colonized asymptomatically by *S. aureus* [25]. *S. aureus* is known to colonize the nasal passage, skin and mucosal surfaces while S. epidermidis is a prevalent colonizer of the skin [26].

In addition to asymptotic colonization *S. aureus* is also known as an opportunistic pathogen that frequently causes wound and skin infections as well as life threatening conditions like pneumonia, sepsis and endocarditis [26–28]. According to Deurenberg *et al* the majority of nosocomial infections today are caused by *S. aureus* [29].

Such infections are especially problematic when caused by methicillin-resistant *S. aureus* (MRSA). In recent years, the spread of MRSA has increased greatly in hospital environments, which is a substantial threat to immunocompromised patients. In addition to hospital-acquired MRSA (HA-MRSA) there are also incidents of community acquired MRSA (CA-MRSA), which signifies MRSA strains that originate from non-hospital environments. CA-MRSA can still spread in hospitals once introduced. CA-MRSA often has additional virulence factors compared to HA-MRSA, e.g. Panton-Valentine-Leukocidin (PVL) [29]. CA-MRSA is regarded as a particular health-threat because of its ability to infect young healthy people who lack the known risk factors for MRSA, as opposed to HA-MRSA which is prevalently a problem in immunocompromised individuals [26, 30].

Genetically, *S. aureus* has been described as a highly clonal species whose core genome is very conserved. Mobile genetic elements, most of which are of phage origin, are what mainly accounts for the diversity of *S. aureus* strains and not least many of the bacterium's virulence factors [15]. That means that the evolution in *S. aureus* seems to be largely phage-driven. Deghorain *et al* report that the 'accessory genome' may constitute as much as up to 25% of a *S. aureus* genome, making the species highly adaptable [28]. Only two years after the introduction of penicillin, a resistant *S. aureus* strain was detected in 1942 and the same repeated two years after the introduction of methicillin [29], which drastically underlines the speed with which *S. aureus* adapts.

Furthermore, it seems that pathogenic *S. aureus* strains favor the mobilization and atypical genomic integration of phages compared to strains that are purely colonizing. This again emphasizes the role of phage derived mobile genetic elements for the pathogenesis of *S. aureus* [28].

4 Sequencing Technologies

4.1 Second generation sequencing

Second generation sequencing, also referred to as next generation sequencing or massive parallel sequencing, is currently the most commonly used technology to produce sequencing data. The target DNA is hereby sheared into fragments which are then clonally amplified and millions of them sequenced in parallel, hence the name massive parallel sequencing [31]. A scheme of the workflow is depicted in Figure 4.1. DNA targets can differ from small PCR fragments (amplicons) to retro-transcribed cDNA in the case of RNA sequencing to de novo sequencing of full genomes [31]. In the context of this thesis I will mostly speak of whole genome sequencing (WGS) which aims to uncover the full sequence of a target genome. There are three principal providers of second generation sequencing: Illumina, 454 pyro sequencing and Ion Torrent, of which Illumina remains the most widely used. Support for 454 sequencing was stopped in 2015 [32]. Each of them outputs a large amount of short sequenced DNA fragments, called reads, that can later be combined into longer contiguous fragments known as contigs by a process called de novo assembly or mapped to a reference genome [33].

The advantages of second generation sequencing are that it is very affordable and produces a large amount of data. The main drawback is that the read length is very short, on the order of 35 to 700 base pairs [34]. This is caused both by limitations in the sequencing technology and by the fact that DNA has to be fragmented for the amplification step. The re-assembly of reads into genomes afterwards is a non-trivial problem. Though various assembly approaches exist, none of them are perfect and it is often not possible to recover a single, closed genome from the data without performing additional PCR over the contig edges or mapping the reads to a closed reference genome.

4.2 Third generation sequencing

In recent years, a new generation of sequencing technologies, commonly referred to as third generation sequencing or single molecule sequencing, has been developed. The major difference to second generation is that instead of generating enormous libraries of short fragments of DNA, samples are sequenced as single molecules without being fragmented. There are two distinct approaches: Nanopore sequencing, as employed by Oxford Nanopore, and single-molecule

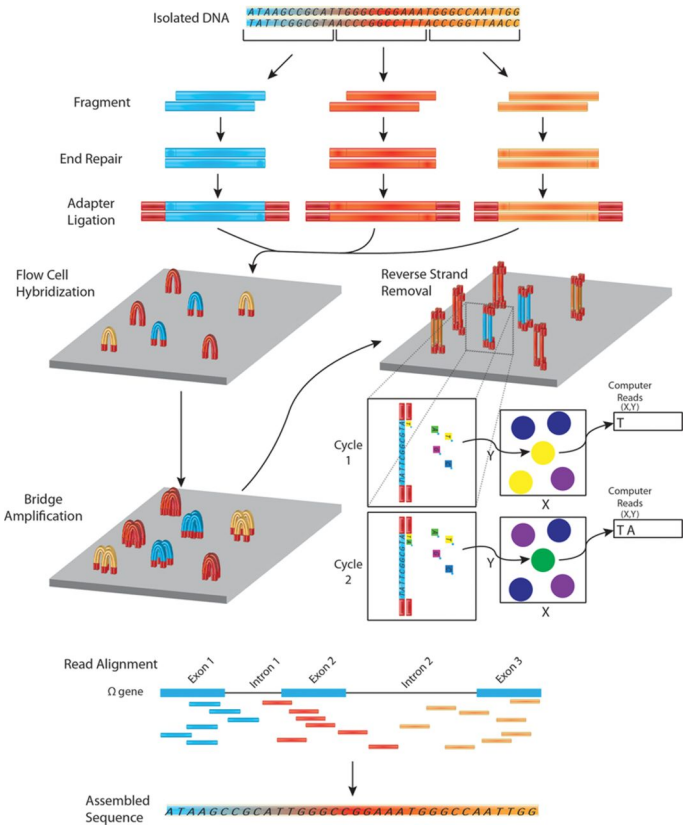


Figure 4.1. Scheme of the Illumina sequencing workflow. The target DNA is first sheared into a fragments and ligated with adapters during a process called library preparation. Afterwards, the DNA fragments are immobilized on the surface of the flow cell and amplified. One of the strands is then removed to prepare for sequencing by synthesis. Specialized nucleotides labeled with fluorescent dye are added. Upon binding, they release a fluorescence signal corresponding to the base that was just added. Reprinted from [35].

real-time sequencing (SMRT), offered by PacBio.

By retaining the target DNA in large fragments during sequencing, possibly encompassing the full genome, the problem of assembly is significantly reduced [36]. Another aspect is a greater ease in sample preparation compared to second generation sequencing, enabling these technologies to be applied outside of laboratory settings [32]. This is especially true for the Oxford Nanopore. Until recently, those technologies were however plagued by high error rates of up to 20% for PacBio [37]. For phages with their very mosaic genomes long read sequencing would be advantageous if the high error rate can be reduced or corrected with short reads from second generation sequencing. For now, this approach remains very expensive.

5 Genomics

5.1 Genomics Tools

The advance of affordable and fast WGS has enabled the development of many sequence based analysis methods.

The first step in the analysis of sequencing data is quality control and trimming of the raw reads. This is necessary because sequencing is not error-free and low quality reads may negatively affect the analysis by introducing noise. There are many different trimming tools available. In my work I have used fastQC [38] for read quality control and PRINSEQ [39] for trimming. Depending on the desired analysis, read data can then be assembled into contigs. Different approaches for assembly exist but the most successful assemblers to date are based on de Bruijn graphs. A de Brujin graph is a graph representation of a sequence (or several sequences) where each k-mer is a node and each edge is an overlap between k-mers. A k-mer is a short sequence fragment of length k. Assembly is then performed by resolving the de Bruijn graph. Examples for de Bruijn graph assemblers are velvet [40] and SPAdes [41], both of which have been used in this thesis.

Other than assembling reads one can also map them to a reference genome or to already assembled contigs. Mapping could in principle be performed with any alignment algorithm but because of the large number of reads, typically in the millions or billions, there are specialized tools for this purpose. The tool used for mapping in this thesis was the Burrows-Wheeler aligner(bwa) [42].

A substantial part of genomics is based on sequence comparison, which can be done either by alignment or based on matching k-mers. The oldest and most widely known alignment algorithm is the basic local alignment search tool (BLAST) [43], which now exists in many variations and has played a pivotal role in the development of the genomics field. BLAST is most commonly used for database searches, such as in the ResFinder [44] and VirulenceFinder [45] tools which employ BLAST to scan a query sequence for known antimicrobial resistance and virulence genes respectively. Another application of BLAST is to estimate distance between sequences via the average nucleotide identity (ANI). ANI is for example be used for species delineation.

Other sequence comparison methods are based on counts of shared k-mers. Two such tools used in this thesis are KmerFinder [46], an algorithm that com-

putes sequence similarity in k-mer space, and cd-hit [47], a sequence clustering algorithm used for homology reduction in datasets.

More specialized tasks often combine sequence similarity search and sequence features such as GC content, tetranucleotide frequencies, genomic signatures such as ribosome binding sites, secondary structure elements ect. Examples used in this thesis are gene calling and functional annotation using prodigal [ref], GeneMarkS [48] and RAST [49].

5.2 Metagenomics

Metagenomics is the sequencing and subsequent analysis of mixed DNA samples, i.e. samples that contain DNA from many different microorganisms without separating those organisms before. Those samples are usually environmental [33].

The shift from single organism genomics to metagenomics is hugely motivated by the desire to understand the communities in which microorganisms live and function as opposed to studying them as isolated entities which is not their natural state [50]. Further, the majority of bacteria are not easily cultivated. The same applies to phages, who naturally exist in close interaction with their bacterial hosts as well as with each other via competition as well as exchange of genetic material during co-infection. As such, the metagenomics approach is well suited to study phage communities in the natural environments. The majority of genomics tools described in the section above are also applicable to metagenomics datasets.

6 Machine learning

Machine learning can broadly be divided into supervised and unsupervised learning tasks. In a supervised learning task both input and output are known and the desired outcome is to find a function that describes their relationship. In contrast to that, when data without known outcomes is available you have an unsupervised learning task. The goal is then to infer underlying principles in the data. In this thesis, only supervised learning was used.

6.1 Generalized linear models

There are many different algorithms that can be used for mapping the input onto the output, however, in this thesis I will focus on generalized linear models. The generalized linear model (GLM) concept unifies several often used statistical models such as linear regression, logistic regression and multinomial regression. In a GLM, model and output are related via a so-called link function. This link function can be understood as determining the type of regression [51].

In this thesis, a logistic GLM was used to model the phage susceptibility of a set of bacterial strains as a function of their present gene families. Logistic regression is the appropriate model type to use for categorical outcome variables and the link function to use is then the logit function. Specifically, the model structure was:

$$y \sim \sum_{i=1}^{N} w_i \cdot x_i \quad with \quad x \in \{0, 1\}$$

where x_i was 1 if the gene family i was present and 0 if it was absent, w_i was the weight assigned to gene family i and y was the predicted susceptibility with 1 being susceptible and 0 being resistant. For details see the publication included in Chapter 9.

6.2 Model training and performance evaluation

Generally in supervised learning tasks, models are trained on training data and then evaluated on testing data. During training the goal is to minimize an error

function between the prediction and the known true result. An often used error function is the mean square error (MSE):

$$MSE = \frac{1}{N} \cdot \sum_{i=1}^{N} (O_i - t_i)^2$$

where N is the number of observations, O_i is the ith predicted value and t_i is the ith true outcome.

However, because of noise inherent in real-world data a maximally low training error does not necessarily correspond to a good model since the model may then start matching the noise instead of an underlying trend. This phenomenon is known as overfitting and is particularly a problem when the feature space is large compared to the number of observations, as typically occurs in high-dimensional models. Overfitting is problematic because the resulting model will be a suboptimal description of the underlying process and hence generalize poorly to the independent evaluation data. Moreover, it will lead to a vast overestimation of the model's performance.

The way to accurately measure model performance is to perform training and testing inside a cross validation (CV) framework [52]. In this framework, the data is firstly divided into partitions, then all but one of the partitions are used to train a model and the last one is used to evaluate it. Each division of partitions into training and test set is called a fold. This process is repeated until each partition in turn has been the testing set. The point of cross validation is to test the model's performance on new, unseen data (i.e. the test set) and thereby get a better estimate of the model's ability to generalize. For this thesis, training and testing of the logistic regression model was performed inside a five-fold cross validation setup.

Another problem present in this dataset, but also in machine learning in general, is data-redundancy. When data points are shared between training and testing set, the classification problem becomes very easy and the model will not learn to generalize to new data. In addition to that the model performance will be overestimated. It is therefore important that data points assigned to different cross validation partitions should not be similar.

There exist different measures of model performance. For a classification task, the receiver operating characteristic (ROC) curve is a good choice as is illustrates the relationship between sensitivity and specificity. The sensitivity, also known as true positive rate, is plotted on the y-axis and 1 - specificity, also known as the false positive rate, is plotted on the x-axis shows. When performing a classification task, the model output is not a binary but a continuous variable. This prediction score is then discretized into a class prediction based

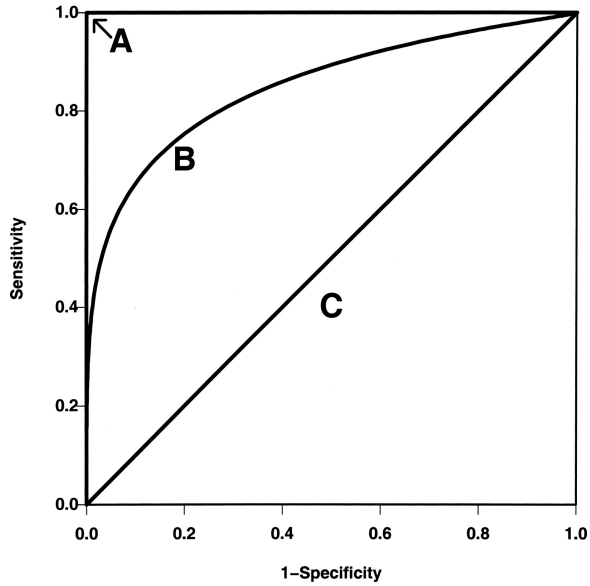


Figure 6.1. Scheme illustrating three different hypothetical ROC curves. A: Perfect performance. In a model with this AUC, it would be possible to place the classification threshold so that all true positives are reported but none of the false positives. B: Realistic performance. A model with this AUC will report more true positives than false positives. C: Random performance. A model with this AUC would report true positives and false positives in equal amounts. Reprinted from [53].

on a classification threshold. Conceptually, a ROC curves displays for every possible classification threshold the ratio of true positives to false positives. To quantify the goodness of a ROC curve one calculates the area under the curve (AUC). A perfect performance would yield an AUC of 1, a random performance an AUC of 0.5. The AUC was used as the measure of performance in the the third study of this thesis, see Chapter 9. An example of three theoretical ROC curves and their corresponding AUCs can be seen in Figure 6.1. One can also calculate separate AUC values for each cross validation fold. If the model is robust, the performance values should be similar across all folds.

6.3 Ridge regression

The model used in the third publication of this thesis was further fitted via Ridge regression during training. Ridge regression is a type of parameter regularization applied during training where the error is penalized with the sum

of squared coefficient weights, also called the L2 norm [51]. The error function then becomes:

$$E = \frac{1}{N} \cdot \sum_{i=1}^{N} (O_i - t_i)^2 + \lambda \cdot \sum_{l=1}^{M} {w_l}^2$$

where N is the number of observations and M is the number of features. Further, O_i is the ith predicted value, t_i is the ith true outcome, w_l is the weight of the lth feature and λ is the strength of penalty.

A ridge regression shrinks the weights of features that have a low importance while maintaining the values of weights that do have general importance [54]. In that way it reduces overfitting. λ is typically tuned to achieve an optimal regularization, as was also done in this thesis. This should be done inside a nested cross validation as depicted in Figure 6.2.

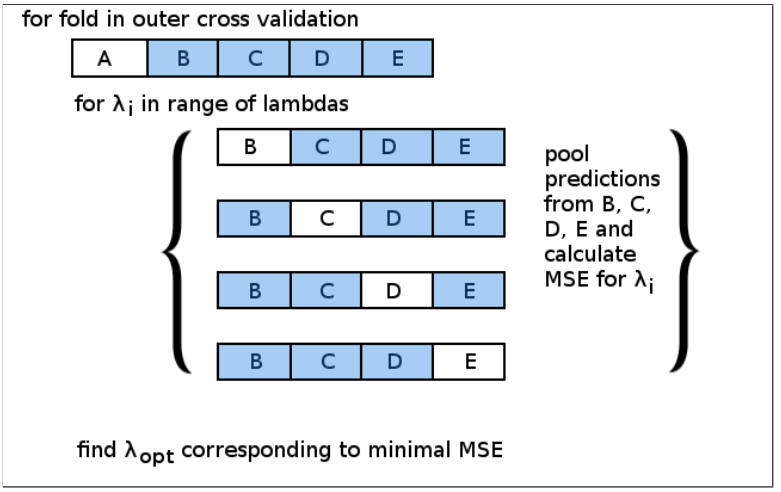


Figure 6.2. Scheme of nested cross validation for finding the optimal strength of penalty λ. Briefly, for each fold in the outer cross validation and for each λ in a range of values, an inner cross validation is performed. In this way, one optimal λ is identified for each outer cross validation fold.

For each outer cross validation fold, an inner cross validation is performed. Note that the inner cross validation only has access to the data in the training set of the corresponding outer cross validation fold. In the inner cross validation, again all but one partition are combined into the inner training set and the remaining partition is used for testing. In order to find the optimal

strength of penalty λ, training and testing are performed for a range of λ values for each inner cross validation fold. In this thesis we chose $1e^{-10}$ to $1e^{5}$. Afterwards, predictions are pooled across all inner cross validation folds (but still separated by λ values). and one mean square error per λ is calculated. This error can be plotted against λ to visualize λ's influence on the model test performance. An example is shown in Figure 6.3. The optimal lambda for the current outer cross validation fold is the one that results in the minimum MSE. This optimal lambda is used to train an additional model using the entire training set of the outer fold and evaluating on the test set of the outer fold.

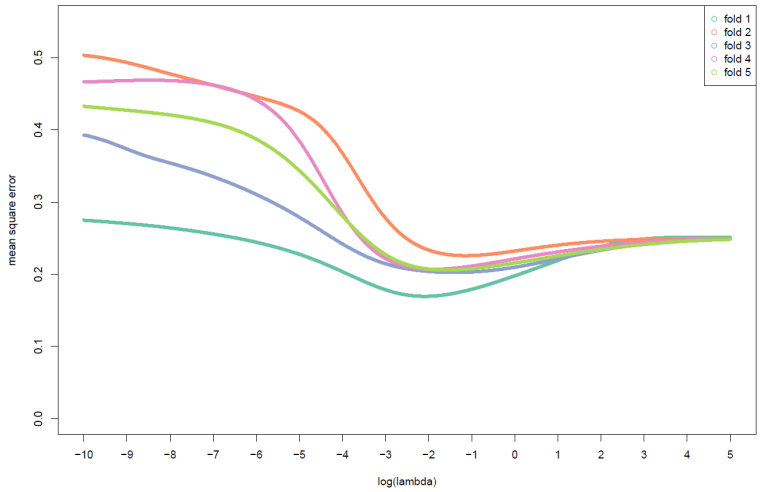


Figure 6.3. Plots of mean square error versus different strengths of penalty λ. A low penalty values the error is comparatively high since there is little regularization. With increasing λ the error generally reduces until it reaches a minimum. Afterwards, the error rises again as strength of penalty becomes too high. It can be seen that the curves for the five different partitions follow a similar trend and their minima coincide to a reasonable degree. This indicates the model is robust. Taken from the supplement of paper 'Host-genomic determinants of phage susceptibility in MRSA', see Chapter 9.

This process is repeated five times for the five outer cross validation folds. A robust model should have comparable optimal lambda values.

6.4 Feature selection

For the dataset used in this thesis the number of features, i.e. gene families, was much greater than the number of observations, i.e. strains. This makes it difficult to find the proper weights and also we can assume that not all features are equally important for the outcome [55]. It is therefore essential to perform feature selection. Feature selection is a process that seeks to limit a model's feature space to only the most important features. This can be done in several ways. To avoid overfitting it is however vital to perform feature selection inside the cross validation framework. Otherwise, information from the test set can influence which features are picked and the test performance will therefore not be an unbiased estimation of performance anymore. For this reason, only information from the training set can used to select features.

In this thesis, feature selection was performed by a pre-selection step followed by a two-step model. During the pre-selection, gene families were filtered based on their p-values resulting from an association analysis between occurrence of the gene family and susceptibility outcome. Since this is done inside the outer cross validation, only data from the respective training set was used in the respective association analysis. Gene families passing the p-value threshold were admitted to an initial regression model. This model was then trained on the same training data used to select the gene families and tested on the left-out test set. Each gene family was assigned a regression weight w_i during training. After that, we moved on to the next outer CV fold and so on five times. In this way, five weights were obtained for each gene family. If a gene family was not picked by pre-selection in one of the folds, its weight in this fold was 'not applicable' (NA). Lastly, from this we selected gene families with regression weights greater than a certain threshold in at least three of the five partitions, to use as features in a final model.

Part II

Studies included in the thesis

7 Sequencing of the INTESTI phage cocktail

The field of phage therapy in general has gone through a great revival in Western research during recent years, owed in large parts to the looming antibiotic resistance crisis. This has understandably created an interest in the phage cocktails already in use in Russia and Georgia. In both of these countries phage therapy has a long history going back to the 1920's/30's and especially the Eliava Institute in Georgia has been a major player in piloting phage research and exporting phages across the Soviet Union [56].

In the following paper, we show what can be done with a metagenomics approach to characterize an existing phage cocktail. This project started as my Master's thesis back in 2014 and then continued on to become the first paper of my PhD. It began with Karina Sreseli, a Georgian secretary in our department, who shared the story of how she had been treated with phage cocktail as a child. At that time my supervisor Mette Larsen was becoming very interested in phages and it was our luck that Karina still had contacts to Georgia - specifically to the Eliava Institute where phage cocktails have been produced since the 1930's. Mette obtained a sample of INTESTI, one of the most famous and longest used phage cocktails of Eliava, and from there on a long journey started during which we met many people all the way from Georgia to the Evergreen State College in Washington, U.S. Some of them become co-authors, others advised us and it ended in a publication that has generated a fair share of interest in the field due to the historical significance of INTESTI phage cocktail.

viruses

MDPI

Article

What Can We Learn from a Metagenomic Analysis of a Georgian Bacteriophage Cocktail?

Henrike Zschach [1], Katrine G. Joensen [2], Barbara Lindhard [1], Ole Lund [1], Marina Goderdzishvili [3], Irina Chkonia [3], Guliko Jgenti [3], Nino Kvatadze [3], Zemphira Alavidze [4], Elizabeth M. Kutter [5], Henrik Hasman [1] and Mette V. Larsen [1,*]

Received: 30 September 2015; Accepted: 30 November 2015; Published: 12 December 2015
Academic Editors: Abram Aertsen and Rob Lavigne

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; henrike@cbs.dtu.dk (H.Z.); b.lindhard@live.dk (B.L.); lund@cbs.dtu.dk (O.L.); henh@ssi.dk (H.H.)
2 National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; kagjo@food.dtu.dk
3 Eliava Institute of Bacteriophages, Microbiology and Virology, 3 Gotua Str., Tbilisi 0160, Georgia; mgoderdzishvili@gtu.ge (M.G.); irinachkonia@yahoo.com (I.C.); gulikojgenti@yahoo.com (G.J.); ninqvatadze@yahoo.com (N.K.)
4 Eliava Biopreparations LTD, 3 Gotua Str., Tbilisi 0160, Georgia; z.i.alavidze@gmail.com
5 Lab 1, The Evergreen State College, Olympia, WA 98505, USA; KutterB@evergreen.edu
* Correspondence: metteb@cbs.dtu.dk; Tel.: +45-45-25-24-25

Abstract: Phage therapy, a practice widespread in Eastern Europe, has untapped potential in the combat against antibiotic-resistant bacterial infections. However, technology transfer to Western medicine is proving challenging. Bioinformatics analysis could help to facilitate this endeavor. In the present study, the Intesti phage cocktail, a key commercial product of the Eliava Institute, Georgia, has been tested on a selection of bacterial strains, sequenced as a metagenomic sample, *de novo* assembled and analyzed by bioinformatics methods. Furthermore, eight bacterial host strains were infected with the cocktail and the resulting lysates sequenced and compared to the unamplified cocktail. The analysis identified 23 major phage clusters in different abundances in the cocktail, among those clusters related to the ICTV genera *T4likevirus*, *T5likevirus*, *T7likevirus*, *Chilikevirus* and *Twortlikevirus*, as well as a cluster that was quite distant to the database sequences and a novel *Proteus* phage cluster. Examination of the depth of coverage showed the clusters to have different abundances within the cocktail. The cocktail was found to be composed primarily of *Myoviridae* (35%) and *Siphoviridae* (32%), with *Podoviridae* being a minority (15%). No undesirable genes were found.

Keywords: phage therapy; Eliava Intestiphage; whole genome sequence analysis; metagenomics

1. Introduction

Antibiotic resistance in human pathogenic bacteria is a threat to public health that has grown immensely in the last years. The World Health Organization (WHO) recognized the severity of the problem in two reports made public in 2012 and 2014, stating that "A post-antibiotic era—in which common infections and minor injuries can kill—far from being an apocalyptic fantasy, is instead a very real possibility for the 21st Century" [1]. It is therefore all the more urgent to secure alternative treatment strategies. Phage therapy is one of the alternatives to antibiotics that for a long time has been underexplored in Western medicine. Bacteriophages, viruses of bacteria, have been employed to combat bacterial infections in certain Eastern European countries since the mid-1920s [2,3]. With

the number of phages on earth estimated at 10^{31} in total [4], they are the most abundant entity in the biosphere and, as natural predators of bacteria, they hold largely untapped therapeutic potential [5].

During the Soviet era, antibiotics were not readily available in the USSR, which contributed to the widespread use of phages for treatment of various sorts of bacterial infections [6]. In particular, the George Eliava Institute in Tbilisi, Georgia, founded in 1923, has more than 90 years of experience in employing phages for treatment of bacterial infections in humans, either as single preparations or in mixtures, *i.e.*, phage cocktails.

Phage therapy is largely regarded as safe and effective in those countries where it is still practiced [7–10]. This is reinforced by the long-standing tradition of its use. The enormous body of experience with clinical phage therapy, which has primarily been reported in non-English languages [11], is now more and more being made available to the scientific community thanks to the concerted efforts of Elizabeth Kutter, Jan Borysowski, Harald Brüssow, Ryszard Międzybrodzki, Andrzej Górski, Beata Weber-Dąbrowska, Mzia Kutateladze, Zemphira Alavidze, Marina Goderdzishvili, Revaz Adamia and others [8,9].

Additionally, a number of more recent trials have been carried out in accordance to the strict guidelines demanded by legislative bodies and published, notably two T4 oral application safety trials [12,13], a trial of *Pseudomonas aeruginosa* phages for treatment of chronic otitis [14], a phase I trial of phage therapy for venous leg ulcers [15] and a trial of Russian phage cocktail administration in healthy individuals [16].

Despite the growing body of evidence on the safety and efficacy of phage therapy, the technology proves hard to transfer despite considerable interest by Western researchers. One of the challenges is a lack of definition and characterization of the phages used, as the exact composition of phages in the cocktails produced in Eastern Europe is largely unknown [17]. Advances in metagenomics and decreasing sequencing costs have made it possible to analyze mixed phage samples without the need to separate the component phages. This is especially essential when the specific bacterial hostsstrains are unknown and the phages can thus not be individually propagated for traditional analysis. This metagenomic approach was first used for marine viral communities in 2002 [18]. One of the latest milestones in this endeavor consists of a metagenomic study of a Russian phage cocktail as well as a safety trial, performed by McCallin *et al.* in 2013 [16].

Here, we present a metagenomic analysis of the longest-used such commercial phage cocktail in the world, still routinely employed for human therapy in the Republic of Georgia. Intesti bacteriophage was created at the Pasteur Institute, Paris by Felix d'Herelle [19] as a multi-component treatment and prophylaxis of intestinal infections. From early on, the preparation is a combination of phage active against *Shigella*, *Escherichia*, *Salmonella*, *Enterococcus*, *Staphylococcus*, *Streptococcus* and *Pseudomonas*. Its advantages lie in its activity against a wide variety of enteric bacteria, allowing it to be used empirically during the first days of gastrointestinal illness, before the microbiological culture results are in, along with its frequent ability to help restore balance to the gut microbiome even where no explicit pathogen has been identified as the cause of the problem.

Intesti bacteriophage was first used clinically in Georgia in 1937 by S. Mikeladze [20]. Already in 1938, M.N. Luria used Intesti-bacteriophage to study 219 patients suffering from either dysentery (84 children and 27 adults with *Shigellashiga* (now known as *Shigella dysenteriae*) or *flexneri*) orhemolytic intestinal disease caused by an unidentified bacterium (54 children and 54 adults).Most had previously been treated unsuccessfully in other ways, but other treatments were stopped during administration of the phage therapy. Adults were given 10 mL and children 2.5–5 mL orally with carbonated water once a day, before meals. Improvement was observed in 163 cases within 1–3 days. The results of this study and a number of others have been summarized in great detail by Chanishvili [21] in her extensive 2009 literature review of the early practical application of bacteriophage research, previously largely available only in Georgian.

There is an unknown, quite large total number of phages in the Eliava Intestiphage cocktail, which has continually been evolved to meet current needs since it was first developed by d'erelle at

the Pasteur Institute. At least one proprietary mother phage stock has been maintained through the years for the phages targeting each genus of bacteria, and each of these is grown separately using a proprietary group of bacterial strains of that genus, which is updated regularly as needed to be able to better target new problem strains that have arisen. Each component thus produced for a new commercial batch is tested on each member of a separate continually-updated broad proprietary group of strains and remade if it does not adequately meet the established high host range for that genus. New phages are periodically added to improve the needed host range for this broadly-applicable commercial cocktail, which has been shown to have such high efficacy in a variety of situations, both as a probiotic and to treat a wide range of gut problems that are often intransigent to more narrowly targeted phage treatments and/or to antibiotic treatment. This challenges most current common regulatory practices in countries other than Georgia, where the carefully defined method of testing and regulation of Intestiphage takes this into consideration, with close cooperation between the Ministry of Health regulatory body and the production facilities. The procedure described above for preparing therapeutic bacteriophage is similar to the procedure described in a chapter on phage production by Felix d'Herelle. The original chapter has been translated into English by Sarah Kuhl and Hubert Mazure [22].

The Eliava Pyophage cocktail, for purulent infections involving *Streptococcus* sp., *Proteus* sp., *Escherichia coli*, *Pseudomonas aeruginosa* and *Staphylococcus aureus*, is the one other cocktail that has evolved in similar fashion over the years. It should be kept in mind that Intestiphage and Pyophage are generic names; other companies in both Georgia and Russia have been making and marketing their own versions for the last couple of decades which have been evolved from the same initial cocktails brought to what is now the Eliava Institute by d'Herelle and are regulated and regularly upgraded in similar fashion. These other versions can be expected to work better in some specific situations, worse in others, depending on their precise composition of phages and of the proprietary hosts that are used in their production and testing. It will be very interesting to also do metagenomic analyses of those other versions and see how their current composition compares, in reflection of this evolutionary process.

2. Materials and Methods

2.1. The Intesti Phage Cocktail

Commercial "Intesti bacteriophage", which is used mainly to treat bacterial infections of the intestine, urinary tract and oral cavity in humans, was kindly provided by Nikoloz Nikolaishvili, director of Eliava Bio Preparations LLC at the George Elivia Institute, Tbilisi, Georgia. The current Eliava Intestibacteriophage contains sterile phage lysates active against *Shigella* (*flexneri*, *sonnei*, Newcastle), *Salmonella* (Paratyphi A, Paratyphi B, Typhumurium, Enteritidis, Cholerasuis, Oranienburg), *Escherichia coli*, *Proteus vulgaris and mirabilis*, *Stapylococcus aureus*, *Pseudomonas aeruginosa* and *Enterococcus*. Intestibacteriophage is used for treatment and prophylaxis of the following bacterial intestinal infections caused by the above mentioned microorganisms: dysentery, salmonellosis, dyspepsia, colitis, enterocolitis, and dysbacteriosis (bacterial overgrowth). Intestibacteriophage treatment per os (via oral route) is used from the first day of disease and is continued for 5–6 days. Intestibacteriophage can be used for prophylaxisin situations where there are large groups of people (for example military or schools), during seasonal peaks in order to reduce occurrence of intestinal infections. The phage preparation developed for therapeutic and prophylactic uses by G. Eliava Institute of Bacteriophages, Microbiology and Virology was awarded in 1978 Gold Medals at the Exhibitions of All-Union National Achievements in Science and Technology.

From the mode of preparation, it follows that the Intesti cocktail is a complex mixture of phages in different abundances, many of which may be closely related. This poses certain challenges both in the sequencing and assembly. Furthermore, different batches of the cocktail may not be identical. Our sample was manufactured in July 2013 and has the batch number M2-501.

2.2. Host-Amplified Samples

In addition to sequencing the complete cocktail as a metagenome, we also amplified the component phages on eight different hosts and isolated DNA from the resulting lysates, which are assumed to be enriched only in the phages capable of infecting the given host. Those samples are therefore reduced in complexity in comparison to the cocktail. The host strains used are part of an in-house Danish collection and listed in Table 1 (Results Section). For each host, 5 mL liquid LB were inoculated with 50 µL from an overnight culture and grown with shaking incubation at 37 °C. After 3 h the day culture was divided into two 2.5 mL samples, of which one was infected with 300 µL of the cocktail and incubated for another 4 h with shaking. When the infected sample had visibly cleared compared to the non-infected sample, indicating that host lysis had occurred, the lysate was filtered through 0.22 µm syringe filters and subsequently treated the same as the Intesti whole cocktail sample (see Sample Preparation). It should be noted that the bacterial host strains used to produce the cocktail in Georgia are proprietary and thus were not available to us in Denmark.

Table 1. List of the strains used to specifically amplify phages from the Intesti cocktail and the number of reads obtained in their sequencing. All strains were tested for susceptibility to the cocktail prior to selection.

Host Bacterial Strain	Number of Reads
Escherichia coli ATCC 25922	358,914
Enterococcus faecalis ATCC 29212	134,966
Pseudomonasaeruginosa 0407431-2	184,790
Pseudomonasaeruginosa PAO1_seq	265,772
Proteus vulgaris CCUG 36761 (ATCC 13315)	64,852
Salmonella typhimurium ATCC 14028	133,980
Shigellaflexneri iran_1s	225,664
Shigellasonnei iran_2s	401,722

2.3. Sample Preparation

All phage samples intended for sequencing were treated with 10 µL (20 units) of 2000 units/mL DNAse (New England BioLabs, Ipswich, MA, USA) per mL of phage lysate and 5 µL of 20mg/mL RNase (Invitrogen, Carlsbad, CA, USA) per mL of phage lysate to remove possible bacterial DNA leftovers. Subsequently, the samples were treated with 4µL of 20 mg/mL Proteinase K (Merck Milipore, Hellerup, Denmark) per mL of phage lysate to open phage capsids, followed by standard DNA extraction by spin column using the Phage DNA isolation kit by NorgenBiotek (Product #46700, Thorold, ON, Canada).

2.4. Sequencing and Genome Assembly

For each sample a DNA library was prepared from 10 ng of sample DNA using the Nextera XT Sequencing kit (Part #15031942, Illumina, San Diego, CA, USA) and sequencing was performed on the Illumina MiSeq system (Illumina, San Diego, CA, USA). The platform's maximum read length was 251 bp corresponding to 251 cycles. The quality of the raw sequencing data was analyzed with the fastQC tool [23] and it was trimmed extensively using the PRINSEQ [24] tool (trimming parameters may be found in the Supplementary Table S1). Following quality trimming, the data were assembled into contigs using the genovo algorithm [25] for the whole cocktail and samples amplified on *E. coli*, *Enterococcus*, *P. aeruginosa* PAO1_seq, *Salmonella*, *Shigellaflexneri* and *Shigellasonnei* and the velvet [26] assembler for samples amplified on *P. aeruginosa* 0407431-2 and *Proteus*.

2.5. Construction of Phage Clusters

Phage clusters were constructed by grouping contigs by their profiles of BLAST [27] hits to NCBI's non-redundant nucleotide collection (October 2014). Those hit profiles were obtained by

applying a quality cutoff on the query coverage of 20% and on the *E*-value of 1×10^{-10} to the raw BLAST results. Contigs were sorted by size and the largest was automatically assigned to the first contig group. Succeeding contigs either joined an existing group or initiated a new one depending on the distance score (see below) between the current contig's hit profile and the group's hit profile. The process is illustrated in Figure 1. Because of the high complexity of the cocktail, we find it useful to think of those drafts as representing clusters of related phages and they are henceforth referred to as clusters.

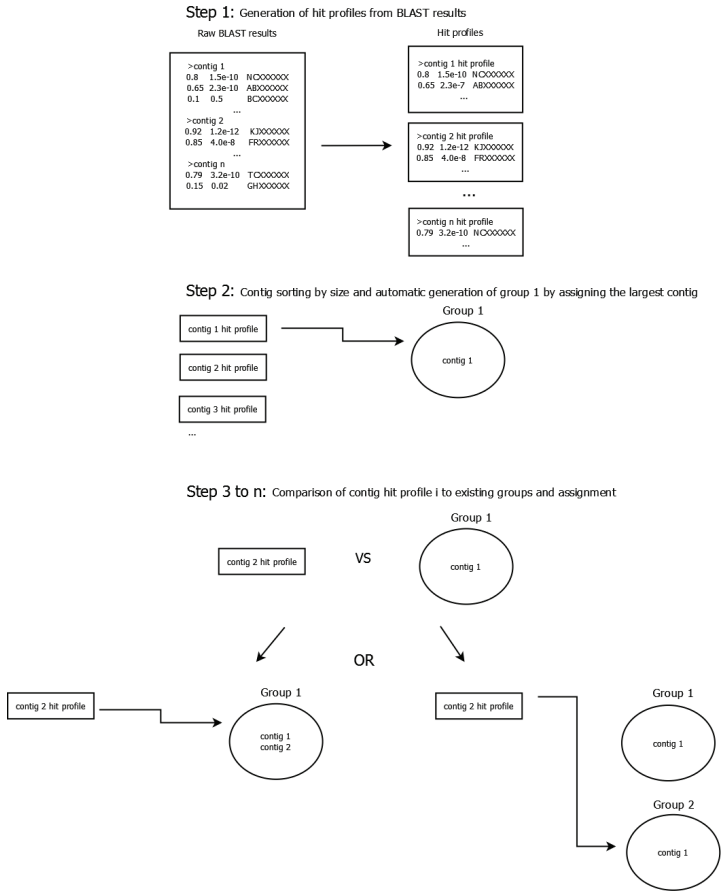


Figure 1. Schematic illustrating the contig grouping process. In a first step, a BLAST search against the non-redundant nucleotide collection is performed for all contigs. Afterwards, a hit profile is generated for each contig by applying a cutoff of 20% on the query coverage and 1×10^{-10} on the *E*-value to the raw BLAST results. During the second step contigs are sorted by size and the largest contig is automatically assigned to group 1. The third step consists of comparing the second-largest contig to all existing groups using the scoring system described in the text and either assigning the contig to the group with the lowest distance score or opening a new group if the lowest score is greater than 0.9. It is repeated until all contigs have been assigned (though some contigs may be the only member of their respective group).

The distance score S_d between two profiles was defined as the average distance of each hit in both profiles such that:

- If the hit is only present in one of the profiles, its distance is 1.0.
- If the hit is present in both profiles, the hit's distance is the absolute value of the difference between the query coverage values, as defined below:

$$S_d\left(profile_l, profile_k\right) = \frac{\sum_{i=1}^{n} \begin{cases} abs\left(querycoverage_{hitiinprofilel} - querycoverage_{hitiinprofilek}\right) \\ 1.0; if\, nothit_i \in profile_l \vee nothit_i \in profile_k \end{cases}}{n} \quad (1)$$

where n is the unique number of hits in profiles l and k.

A contig group's hit profile is the weighted average of the hit profiles of its member contigs and it was updated every time a contig joined the group. The query coverage, *i.e.*, to which extent a contig is covered by that particular hit was thereby used as a scaling property ranging between 0 and 1. The more of a contig is represented by the hit, the bigger the influence of that hit on the difference score. This was done to address the modular nature of phage genomes [28].Contigs that had database hits which were not shared by any other contigs were compared to known phages with regard to length, coverage of the contig by the reference and percent sequence identity, in order to establish whether they could be representing full phage genomes. Contig groups smaller than 5 kb in total size were excluded from further analysis. They represent less than 1% of the assembly size and mostly had hits to bacterial DNA, though upon further investigation many of those hits turned out to be confirmed or suspected prophage or mobile element regions.

We further employed BLAST to identify contig groups from different samples that are thought to originate from the same phage cluster. Contigs from the sample amplified on a *Proteus* host were compared to NCBI's non-redundant nucleotide collection (October 2014) and after checking for sufficiently high depth of coverage those without hits were considered as belonging to novel *Proteus* phages.

2.6. Analysis of the Depth of Coverage

The average depth of coverage was calculated for each contig by mapping the reads that were previously used for assembly back to the contig. Following that, the average depth of coverage for each cluster was calculated from the depth of coverage of its member contigs. We herein incorporated contig length as a scaling factor in the calculation and thereby obtained the weighted arithmetic mean of the cluster's depth of coverage and weighted standard deviation of the same as defined below.

Depth of coverage of contig i,

$$x_i = \frac{N \times L}{w_i} \quad (2)$$

weighted mean depth of coverage of cluster j

$$\overline{x}_j = \frac{\sum_{i=1}^{n} w_i \times x_i}{\sum_{i=1}^{n} w_i} \quad (3)$$

and weighted standard deviation of the depth of coverage of cluster j

$$\overline{\sigma}_j = \sqrt{\frac{\sum_{i=1}^{n} w_i \times \left(x_i - \overline{x}_j\right)^2}{\sum_{i=1}^{n} w_i}} \quad (4)$$

as used in this study, where N = number of reads mapped to contig i, L = average read length, x_i = depth of coverage of contig i, weight w_i = length of contig i and n = the number of contigs in cluster j.

Mapping was performed using the Burrows-Wheeler Alignment tool (BWA) [29]. Prior to mapping, reads were quality trimmed (specifics may be found in Supplementary Table S1), however, duplicates were not removed as had been done for the assembly.

2.7. Gene Prediction and Functional Annotation

Putative genes were predicted in both grouped and un-grouped contigs. Nineteen near complete draft genomes were submitted to the annotation server RAST [30] for functional annotation. Additionally, gene calling was performed on all contigs using the GeneMarkS algorithm [31], followed by a BLAST search against NCBI's non-redundant protein database to infer annotation from existing homologs and achieve an overview of the functions present in the phage cocktail. Annotation was hereby extracted from the top BLAST hit with the additional requirement that the match to this top hit had an *E*-value smaller than or equal to 1×10^{-10}. The results of the two approaches were then compared. Two genes were considered to be the same if their start and end coordinates were less than 10% of the gene length apart and in frame of each other; that is, if the difference between the coordinates for the two genes was a multiple of three. The obtained annotation was subsequently text-mined for genes considered to be undesirable in phage therapy, such as bacterial virulence factors and genes related to lysogeny [32], as well as for genes speculated to enhance the phages' efficacy. For this part, we chose to focus on methylase genes which have been discussed as a method to evade restriction by the bacterial host [33]. Furthermore, the complete assembly was scanned against a database of known genes for acquired antimicrobial resistance by using the ResFinder tool [34] and against a database of known virulence genes in *E. coli*, *Enterococcus* and *Staphylococcusaureus* using the VirulenceFinder tool [35]. No gene prediction and annotation was performed in the host-amplified samples.

2.8. Host Range Estimation

Lastly, in order to verify the cocktail's capability to cause lysis of the specified pathogens, five to ten strains were selected for each pathogen and tested for susceptibility towards the phage cocktail by streaking the bacteria onto an agar plate perpendicular to a streak of phage solution. The selection was oriented towards maximum diversity, including strains from different geographical origins and different host reservoirs. For the pathogens only listed at genus level, different species were tested. The strains and test results can be found in Supplementary Table S2. If lysis occurred in the intersection zone, the bacterial strain was registered as being susceptible to the cocktail. Ambiguous results were repeated in triplicate.

3. Results

3.1. Sequencing Statistics

After quality trimming the sequencing of the full Intesti cocktail resulted in 440,392 reads with an average read length of 174.9 bp. *De novo* assembly yielded 420 contigs ranging in size from 500 to 134,226 bp and a total assembly size of 2041 kb.

In the host-amplified samples, the sequencing depth varied between the different samples. This is indicated by the differing number of reads, see Table 1. Some of the reasons for this could be a variation in the input DNA concentration, as well as amplification bias during library preparation and during the sequencing process.

Table 2. Overview of selected characteristics of the phage clusters identified in the Intesti sample. If known, the family, subfamily and genus of the closest database reference as specified by the ICTV are given. In some cases, the closest reference phage has not been incorporated into the phage taxonomy yet but other references have. For those, both the closest reference and the closest reference within the taxonomy scheme are given. The genus "rv5-like virus" has been proposed by several authors [36,37], but is not confirmed in the current (2014) ICTV release. Remark that Bacteriophage G1 is annotated as a *Staphylococcus* phage.

Phage Cluster	Cluster Size in bp	Reference Accession	Average Coverage of Phage Cluster	Average Percent Identity	Reference Phage Description Line	Phage Family	Subfamily	Genus	Size RatioCluster/Reference
D1	142,025	KC012913.1	99.97	99.80	*Staphylococcus phage Team1*, complete genome	*Myoviridae*			1.01
		AY954969.1	97.98	99.74	*Bacteriophage G1*, complete genome *		*Spounavirinae*	*Twortlikevirus*	1.02
D2	76,960	JX415536.1	87.89	87.60	*Escherichia phage KBNP135*, complete genome	*Podoviridae*			1.00
D3	87,828	KC862301.1	98.97	96.16	*Pseudomonas phage PAK_P5*, complete genome	*Myoviridae*			1.00
D4	69,023	KF562340.1	87.20	94.02	*Escherichia phage vB_EcoP_PhAPEC7*, complete genome	*Podoviridae*			0.96
D5	150,530	FR775895.2	92.41	98.16	*Enterobacteria phage phi92*, complete genome	*Myoviridae*			1.01
D6	81,563	AB609718.1	35.55	77.46	*Enterococcus phage phiEF24C-P2* , complete genome	*Myoviridae*			0.57
D7	58,193	KJ094032.2	77.23	88.35	*Enterococcus phage VD13*, complete genome	*Siphoviridae*	-	*Sap6likevirus*	1.06
D8	50,277	HM035024.1	98.16	90.67	*Shigella phage Shfl1*, complete genome	*Siphoviridae*	-	*Tunalikevirus*	0.99
D9	39,912	EU734172.1	88.25	93.45	*Enterobacteria phage EcoDS1*, complete genome	*Podoviridae*			1.02
D10	145,982	KJ190158.1	93.95	93.00	*Escherichia phage vB_EcoM_FFH2*, complete genome	*Myoviridae*			1.05
		DQ832317.1	93.72	92.62	*Escherichia coli bacteriophage rv5*, complete sequence		-	*"rv5-like virus"* *	1.06
D11	61,791	JX094499.1	96.33	92.95	*Enterobacteria phage Chi*, complete genome	*Siphoviridae*			1.04
		KC139512.1	95.15	93.86	*Salmonella phage FSL SP-088*, complete genome		-	*Chilikevirus*	1.04
D12	60,451	KJ010489.1	54.57	87.35	*Enterococcus phage IME-EFm1*, complete genome	*Siphoviridae*			1.42
D13	188,630	GU070616.1	88.67	94.90	*Salmonella phage PVP-SE1*, complete genome	*Myoviridae*		*"rv5-like virus"* *	1.29

Table 2. *Cont.*

Phage Cluster	Cluster Size in bp	Reference Accession	Average Coverage of Phage Cluster	Average Percent Identity	Reference Phage Description Line	Phage Family	Subfamily	Genus	Size RatioCluster/Reference
D14	133,015	JX128259.1	94.55	96.24	*Escherichia phage ECML-134*, complete genome	*Myoviridae*			0.80
		DQ904452.1	93.42	96.00	*Bacteriophage RB32*, complete genome		*Tevenvirinae*	*T4likevirus*	0.80
D15	43,967	GQ468526.1	87.06	91.27	*Enterobacteria phage 285P*, complete genome	*Podoviridae*			1.12
		FJ194439.1	87.13	90.61	*Kluyvera phage Kvp1*, complete sequence		*Autographivirinae*	*T7likevirus*	1.11
D16	46,882	KM233151.1	93.68	91.47	*Enterobacteria phage EK99P-1*, complete genome	*Siphoviridae*			1.06
		JX865427.2	91.64	91.03	*Enterobacteria phage JL1*, complete genome			*Hk578likevirus*	1.08
D17	41,098	AY370674.1	88.68	94.28	*Enterobacteria phage K1-5*, complete genome	*Podoviridae*	*Autographivirinae*	*Sp6likevirus*	0.93
D18	41,016	HE775250.1	94.95	91.57	*Salmonella phage vB_SenS-Ent1* complete genome	*Siphoviridae*			0.97
		JX202565.1	92.76	91.41	*Salmonella phage wksl3*, complete genome			*Jerseylikevirus*	0.96
F1	13,855	HG518155.1	99.97	99.02	*Pseudomonas phage TL* complete genome	*Podoviridae*			0.30
		AM910650.1	91.92	97.11	*Pseudomonas phage LUZ24*, complete genome		-	*Luz24likevirus*	0.30
F2	11,476	EU877232.1	99.94	91.42	*Enterobacteria phage WV8*, complete sequence	*Myoviridae*	-	*Felixounalikevirus*	0.13
F3	5706	HQ665011.1	83.42	86.09	*Escherichia phage bV_EcoS_AKFV33*, complete genome	*Siphoviridae*			0.05
		AY543070.1	82.09	87.59	*Bacteriophage T5*, complete genome		-	*T5likevirus*	0.05
F4	2624	EF437941.1	98.59	97.76	*Enterobacteria phage Phi1*, complete genome	*Myoviridae*	*Tevenvirinae*	*T4likevirus*	0.02
Proteus phage	104,213	-	-	-	-	*Siphoviridae*			-

3.2. Recovered Phage Clusters

Within the cocktail, 22 phage clusters were recovered by grouping using BLAST hit profiles (see Materials and Methods); plus one novel *Proteus* phage cluster was cluster identified by comparing contigs without hits between the Intesti sample and the *Proteus* host-amplified sample. All clusters are listed in Table 2. They are denoted by a capital D and numbered, except for four smaller clusters under 30 kb in size, which are regarded as containing fragments of phages and therefore denoted by capital F instead. The reason those four clusters are thought to be fragments is that they are small compared the known phages they resemble most, while the other clusters are of similar or greater size than their BLAST hit. It is acceptable for a cluster to be of greater size since the cluster size is cumulative of all member contigs and there can be several variant phages. Overall, clusters ranged in size from 13.4 to 212 kb and were composed of between one and 56 contigs. Seventy contigs, which together make up 217 kb of sequence or 10.6% of the total assembly size, had no significant hits to NCBI's nr nucleotide database. They could therefore not be assigned to a cluster. A list of clusters recovered in the host-amplified samples may be seen in Supplementary TableS3.

3.2.1. Similarity to Known Phages

The most significant BLAST hits used to form the phage clusters were used to examine which known phages a cluster seems to be related to. In Table 2 the reference phage with the highest identity is listed for each cluster, together with the family and, if given, subfamily and genus of that phage according to the ICTV. In cases where there is no taxonomical data available for the closest match but for another match, this reference phage is also listed (compare D14, D15, D16, D18, F1 and F3). Based on the phage family of their closest references, we inferred the potential family association of the clusters. A BLAST search of the predicted tail fiber, DNA polymerase and capsid genes of the *Proteus* phage revealed them to be most similar to those of *Siphoviridae*. We therefore predict the *Proteus* phage cluster to belong to the *Siphoviridae* and count the reads mapped to it into that family. While larger than most studied *Siphoviridae* (which are around 50 kb), the 104 kb Proteus cluster is still smaller than the genomes of the T5 genus of phages. The depth of coverage is quite even along the two contigs in this cluster, so it seems unlikely that the length has been artificially increased through collapsing multiple phages into the cluster.

The clusters could be divided into three groups based on their similarity to their reference phages: Clusters with several highly similar references (query coverage and percent identity >90%), cluster with medium similar references (query coverage and percent identity between 90% and 70%) and clusters that were very distant from all publically available phage sequences. The clusters with several highly similar references are D1, D3, D8, D10, D11, D14, D16, D18, F1, F2 and F4. Specifically for D1 and D3, the resemblance to their closest database reference was very pronounced. We therefore conclude that we have identified phages that appear to be of the same phage species as *Staphylococcus* phage Team1 (KC012913.1) and *Pseudomonas* phage PAK_P5 (KC862301.1), respectively, in the Intesti phage cocktail. The other eight clusters in these groups can also be viewed as fairly close relatives of the clusters described by their reference phages. The second group of clusters, with a slightly lower but still apparent similarity to their references, was D2, D4, D5, D7, D9, D13, D15, D17 and F3. These clusters contain parts that differed from their references, either because they were acquired from other phage species or because they are novel. In contrast, the references for the clusters D6 and D12 were quite distant, as can be seen by the low query coverage. This means that large parts of those two clusters are novel.

Regarding the inferred taxonomy of the clusters, we were able to assign 13 of the clusters to a suspected genus. Of those, four were assigned to the *Myoviridae* genera *Twortlikevirus*, *T4likevirus* (two clusters) and *Felixounalikevirus*. A further six clusters were assigned to the *Siphoviridae* genera *Sap6likevirus*, *Tunalikevirus*, *Chilikevirus*, *Hk578likevirus*, *Jerseylikevirus* and *T5likevirus*. Finally, three clusters were assigned to the *Podoviridae* genera *T7likevirus*, *Sp6likevirus* and *Luz24likevirus*. Two more clusters had reference phages that have been proposed for the new *Myoviridae* genus *rv5-like*

virus, however this genus remains unconfirmed in the 2014 ICTV release. Another six clusters have reference phages, which have not been placed in the official taxonomy yet. Furthermore, the cluster D6 and the Proteus phage cluster may represent entirely new taxa.

3.2.2. Depth of Coverage in the Intesti Clusters

It was found that the weighted average depth of coverage varied considerably between clusters, indicating a different abundance of those clusters within the cocktail (compare Figure 2). D6 and D12 as well as the *Proteus* phage cluster were found to be particularly abundant with an average depth of coverage greater than 150×. In contrast, the clusters D3, D4, D5, D8, D11, D14, D17 and D18 had a very low average depth of coverage of 10× or less.

Furthermore, we observed that many clusters exhibited some degree of variation in the depth of coverage between their member contigs, evident by the weighted standard deviation, which is shown as error bars in Figure 2. Upon inspection, we found that this was generally caused by a few contigs with a very different depth from the rest (compare supplementary Figure S1). We reason that those contigs can be explained by one of the following two scenarios.

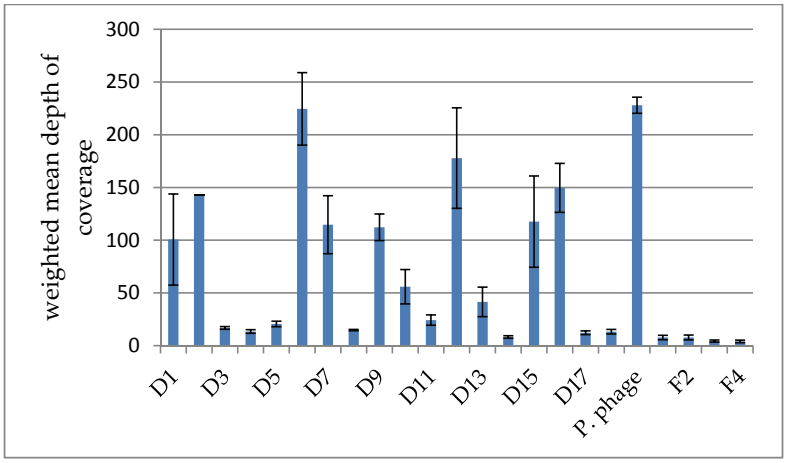


Figure 2. Comparison of the weighted mean of the depth of coverage between clusters in the Intesti sample. The weighted standard deviation is depicted as error bars. Note that cluster D2 is composed of only one contig and the standard deviation is therefore not applicable. It can be seen that the depth varies greatly between clusters, reflecting the different abundances of the represented phage types in the cocktail.

In a sufficiently closely related cluster, most of the common genome will assemble into a few long contigs with a high depth of coverage. The parts that differ between phages in the cluster, however, assemble into contigs that have a much lower depth. In that case, the depth of coverage is proportional to how common the module represented by that contig is within the cluster. Low coverage contigs may also be variants of the more common sequence contained in the high coverage contigs. Contrary to that, in a less closely related cluster, the parts of the phage genome that are shared can assemble into a few chimeric contigs instead of being placed in their respective genomes, causing those contigs to have excess coverage compared to the rest.

Furthermore, we looked at the abundances of the phage families by summing the reads mapped to all clusters inferred to be *Myoviridae*, the same for *Podoviridae* and *Siphoviridae*. Reads mapping to contigs not assigned to a cluster are counted as unknown family. Doing that, we observed 35% *Myoviridae*, 15% *Podoviridae* and 32% *Siphoviridae* in the reads. On top of that, 18% of the total reads

are of unknown family. Observe that those fractions refer to reads that are quality trimmed but not redundancy reduced. When doing the same procedure with redundancy reduced reads, the fractions change to 41% *Myoviridae*, 16% *Podoviridae*, 29% *Siphoviridae* and 14% unknown family.

3.2.3. Depth of Coverage in the Host-Amplified Samples and Comparison of Phage Clusters between Samples

After performing contig grouping in the host-amplified samples, we examined each clusters' highest scoring hits to phage in the non-redundant nucleotide collection and compared to the highest scoring hits in the Intesti clusters. Based on that, we identified clusters across samples that appeared to be synonymous. Using the ratio of the depth of coverage in the host amplified sample to the depth of coverage in the non-amplified Intesti sample, we were able to identify the infecting clusters since those experienced a great rise in coverage, up to 1000-fold (compare Table 3). All of the samples show significant amplification in only a few of the clusters. D14 was able to infect *E. coli* as well as both *Shigella* species, which is concurrent with the notion that those two species are closely related [38]. The two *Shigella* species tested were found to be susceptible to the same two clusters D14 and D15. Both of those appeared to be relatives of *Escherichia* or *Enterobacteria* phages. The *Enterococcus* and *Salmonella* samples shared two infecting clusters, namely D18 and F2. The authors are doubtful of the truth of this result, as *Enterococcus* is Gram positive and *Salmonella* Gram negative. It has therefore been removed.

Table 3. Depth of coverage ratio of host-amplified samples to the Intesti sample. Combinations with a ratio greater than 1.0 are indicated by green background coloring. Those are thought to be the infecting clusters, as they are more abundant in the host-amplified sample than in the original one. In the last line is shown a phage cluster, which has not even been considered in the initial contig grouping of the Intesti sample because of its small size of only 1346bp and low depth of coverage of only 2×. It has, however, been greatly amplified on *P. aeruginosa* strain PAO1.Results regarding the amplification on Salmonella were inconclusive and therefore removed (see text).

Cluster	*E. coli*	*Enterococcus*	*P. aeruginosa* PAO1	*P. aeruginosa* PA0407	*Shigella flexneri*	*Shigella sonnei*	*Proteus*
D1	0.03	0.00	0.00	0.00	0.02	0.00	0.00
D2	0.02	0.00	0.00	0.00	0.02	0.02	0.00
D3	0.30	0.00	0.00	22.29	0.10	0.00	0.00
D4	0.09	0.00	0.00	0.00	0.00	0.00	0.00
D5	0.11	0.00	0.00	0.00	0.00	0.00	0.00
D6	0.06	0.00	0.02	0.00	0.01	0.01	0.00
D7	0.05	2.57	0.00	0.00	0.02	0.00	0.00
D8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D9	0.04	0.00	0.00	0.00	0.02	0.00	0.00
D10	0.08	0.00	0.06	0.00	0.02	0.00	0.00
D11	0.13	0.13	0.00	0.00	0.00	0.00	0.00
D12	0.04	0.00	0.01	0.00	0.02	0.00	0.00
D13	0.05	0.05	0.00	0.00	0.04	0.00	0.00
D14	4.74	0.00	0.00	0.00	2.82	2.06	0.00
D15	0.04	0.00	0.00	0.00	4.97	9.84	0.00
D16	0.04	0.00	0.00	0.00	0.02	0.02	0.00
D17	47.17	0.00	0.00	0.00	0.00	0.00	10.01
D18	0.37	-	0.00	0.00	0.00	0.00	0.00
F1	0.00	0.00	1.47	0.00	0.00	0.00	0.00
F2	0.00	-	0.00	0.00	0.00	0.00	0.00
F3	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F4	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Proteus	0.04	0.00	0.02	0.00	0.00	0.01	0.12
*	0.00	0.00	1044.20	0.00	0.00	0.00	0.00

Note: The cluster marked by an asterisk (*) exists in the Intesti sample but has not been named due to its small size and low depth (see table header).

BLAST-based comparison of those infecting clusters confirmed that they had a highly similar sequence content to the clusters in the unamplified Intesti sample. With the exception of two clusters amplified on *P. aeruginosa* PAO1, all others clusters were also of similar length when compared between samples. F1, which is a fragment cluster in the Intesti sample, probably due to low

abundance of those phages in the cocktail, nearly doubled in size to 22,920 bp on the PAO1 sample. Despite this, about half of the sequence content of the F1 cluster in the Intesti sample is not represented in the F1 cluster in the PAO1 sample. This indicates that F1 contains at least two distinct phages, only one of which was amplified on PAO1, and this amplification enabled us to recover more of the sequence of that phage. Furthermore, a new cluster of length 45,478 bp appeared in the PAO1 sample. There is evidence of this cluster in the Intesti sample but was not treated as such due to its very small size of 2392 bp and low depth of coverage of 1.78×. Those results gave us more confidence that the clusters defined by us are meaningful within the context of the cocktail.

Certain samples as e.g., the one amplified on *E. coli* also contained many different clusters in low abundance. We believe that those phages are un-amplified phages carried over from the cocktail when the host culture was infected. This is backed up by the fact that those clusters are synonymous to Intesti clusters with a high depth of coverage and they are predominantly observed on those host-amplified samples that featured a high read-count. Additionally, we found no indication that the phage cluster we think to be a cluster of *Proteus* phages is capable of infecting the *Proteus vulgaris* strain we used for amplification.

3.2.4. Gene Prediction and Functional Annotation in the Intesti Clusters

Gene prediction via GeneMark S on all contigs yielded a total of 3013 genes, 2577 of which were predicted on the contigs that were assigned to a phage cluster and 258 of which were predicted on unassigned contigs. 2864 genes (95%) had hits to NCBI's non-redundant protein database and annotation was retrieved from the top hits. It was however found to be of limited usefulness since it is not standardized or focused on molecular function and often consists of unspecific terms such as "hypothetical protein" or terms that only carry meaning within the genome they were originally annotated in like "ORF3245".

The RAST service, which was only used on the phage clusters, predicted 2408 genes. RAST uses homology to genes in internal databases to retrieve annotation for the genes it calls. If this fails, the annotation line "hypothetical protein" is given, though it can also be obtained by homology to a gene already annotated in that way. A total of 893 genes (37%) carry the "hypothetical protein" annotation. The overlap between genes predicted by RAST and GeneMarkS was 2230 genes.

Phages with the ability to integrate into the host's genome are known to often carry genes that increase their host's fitness, among those resistance genes and virulence factors. For that reason, integrase genes are generally regarded as undesirable in a phage therapy context [3]. The full assembly of the cocktail's metagenome was scanned against databases of resistance genes and virulence genes using the ResFinder [34] and VirulenceFinder [35] tools. Neither scan detected the presence of any known antimicrobial resistance genes or bacterial virulence factors for *E. coli*, *Enterococcus* or *Staphylococcus*. Text mining the annotation for the terms "resistance" and "virulence" returned seven genes in the RAST annotation, which are listed in Table 4. All but one of those genes were also predicted by GeneMarkS, but differently annotated through BLAST. None of these genes, however, seemed to be related to antibiotic resistance. A literature search determined that the identified resistance genes were related to antiseptic resistance, which is not regarded as problematic as antibiotic resistance [39] but also not desirable, especially in relation to the treatment of pathogens. On the other hand, antiseptics like acridine and acriflavine have been shown to inhibit phage activity [40,41], so the presence of resistance genes against those agents might be a tradeoff between achieving the highest possible safety and retaining efficacy of the phage cocktail. Furthermore, one of the most thoroughly lytic phages T4 can become resistant to inhibition of replication by acridine and acriflavine [42].The two proteins annotated as "Phage virulence-associated protein" have tail proteins among their closest BLAST hit, so it can be assumed that the term refers to virulence of the phage towards its host and not to bacterial virulence factors.

Table 4. List of genes potentially relevant for efficacy, found by text mining annotation results. The annotation column details whether the gene was found in the annotation provided by RAST, by BLAST or both. If only one is named the other method either did not predict the gene or annotated it differently. Top BLAST hit, query coverage as given by BLAST and percent identity as given by BLAST are only filled out if applicable. Most genes which were picked up for their RAST annotation still have a BLAST hit description line, query coverage and percent identity values because that gene was also called by GeneMarkS. In any case, the last two columns apply to the BLAST hit, but not necessarily to the hit in the RAST databases. The acridine resistance gene evidenced in D14 was not called by GeneMarkS. If the gene was picked up for its BLAST annotation column 2 and 5 are identical.

Text Mining Term	Description Line	Part of Cluster	Annotation by	Top BLAST Hit Description Line	Query Coverage	Percent Positives
"virulence"	Phage virulence-associated protein	D1	RAST	ORF002 (Staphylococcus phage G1)	100%	100%
	Phage virulence-associated protein	D6	RAST	putative adsorption associated tail protein (Enterococcus phage phiEF24C)	100%	95%
"resistance"	Acridine resistance	D14	RAST	-	-	-
	Acriflavin resistance protein	D3	RAST	hypothetical protein PAK_P500103 (Pseudomonas phage PAK_P5)	100%	100%
	Tellurium resistance protein TerD	D5	RAST	Phi92_gp172 (Enterobacteria phage phi92)	100%	100%
	Tellurium resistance protein TerD	D5	RAST	Phi92_gp173 (Enterobacteria phage phi92)	100%	100%
	Tellurite resistance protein	D5	RAST	Phi92_gp178 (Enterobacteria phage phi92)	100%	100%
"methyltransferase" or "methylase"	DNA methylase	D7	RAST/BLAST	See "Description line"	100%	99%
	DNA N-6-adenine-methyltransferase	D8	RAST/BLAST	See "Description line"	94%	90%
	putative site specific DNA methylase	D8	BLAST	See "Description line"	100%	99%
	DNA methyltransferase	D13	RAST/BLAST	See "Description line"	100%	99%
	putative DNA N-6-adenine methyltransferase	D10	RAST/BLAST	See "Description line"	100%	99%
	Dam methylase	D8	BLAST	See "Description line"	100%	100%
	putative DNA adenine methylase	D11	BLAST	See "Description line"	100%	100%
	putative DNA methyltransferase	unassigned	BLAST	See "Description line"	100%	100%
	DNA adenine methyltransferase	D14	BLAST	See "Description line"	100%	99%
	putative DNA adenine methylase	D11	RAST/BLAST	See "Description line"	100%	97%
	dCMPhydroxymethylase	D14	RAST/BLAST	See "Description line"	100%	100%
	putative adenine methyltransferase	D10	RAST/BLAST	See "Description line"	100%	98%
	DNA-cytosine methyltransferase	D5	RAST	Phi92_gp043 (Enterobacteria phage phi92)	100%	99%
	Adenine-specific methyltransferase	D5	RAST	Phi92_gp155 (Enterobacteria phage phi92)	100%	99%
"integrase"	Phage integrase	D2	RAST/BLAST	putative integrase (Escherichia phage KBNP1711)	100%	98%
	Phage integrase	D4	RAST/BLAST	integrase (Enterobacter phage IME11)	100%	99%

In addition to that, both annotation methods found two genes described as integrases in the clusters D2 and D4. The D2 integrase had a sequencing coverage of 110×, while the D4 integrase had a sequencing coverage of 11×. Both are congruent with to the coverage of the contigs they are placed in. Furthermore, both genes showed high similarity to known integrase genes (see Table 4). However, no statement can be made about the lysogenic or lytic nature of D2 and D4 phages since the integrity of the lysogeny module was not tested in the lab.

Lastly, 10 genes described as "methyl-transferase" or "methylase" were found in RAST's annotation and 13 in the BLAST based annotation. We speculate that those genes may have a positive influence on efficacy as they can enable the phage to evade restriction-modification based defense systems as was detailed in a review by Samson *et al.* [33].

3.2.5. Evaluation of Sequencing Depth of the Cocktail

A rarefaction curve was made by assembling discreet fractions of the quality trimmed reads and plotting the total assembly size *vs.* the fraction of reads used. The reasoning behind this was that if the phage cocktail has been sequenced sufficiently deeply, the assembly size will converge as more reads will add depth to the existing contigs instead of creating new ones. This behavior was indeed observed (compare Figure 3). It can be seen that the rarefaction curve is not completely flattened out, indicating that there may be rare phages not represented in the reads. Still, we reason that while the sample is not sequenced to its entire diversity we have succeeded in covering the majority of the phages present. Furthermore, when re-mapping reads to the finished assembly, 425,960 (97%) of the 440,392 reads map properly.

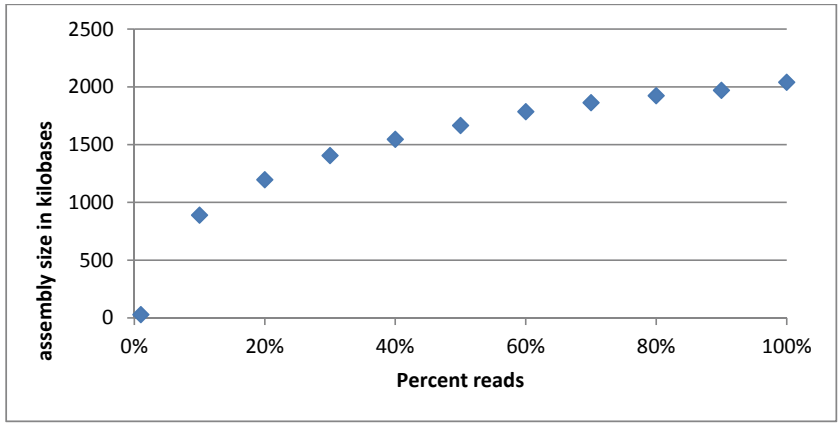


Figure 3. Rarefaction curve of the Intesti sequencing sample. The curve appears to flatten out as the percentage of reads used increases, indicating that the total assembly size is converging. This means that the most common phages are well represented in the sequencing reads. Phages that are in low abundance may not be adequately covered though.

3.3. Host Range Estimation

In a small scale *in vitro* experiment we found the host range of the cocktail to be largely consistent with the specification given by the producer. Five to ten strains were tested for each pathogen listed on the package. The exact number of strains tested and the fraction of strains found susceptible are given in Table 5. The streaking tests confirmed that the cocktail was in principle able to cause lysis of strains of all seven pathogens specified by the producer, albeit with differing specificity for the different pathogens. The apparent low efficiency in lysis of *Staphylococcus* is due to the fact that only

five of the ten tested isolates were *S. aureus*, of which all but one were susceptible. This can be seen in Supplementary Table S2, which also contains a complete list of the specific strains tested.

Table 5. Fraction of the strains found to be susceptible for each pathogen tested. Observe that this is only a small-scale experiment. All strains are part of an in-house collection.

Pathogen	Susceptible Strains
Salmonella Enterica	10/10
Staphylococcus	5/10
Shigella	5/5
Pseudomonas Aeruginosa	5/7
E. coli	2/6
Proteus	3/5
Enterococcus	2/5

4. Discussion

4.1. Completeness and Accuracy of the Analysis

The rarefaction curve showed that the phages that are numerically in the majority appear to be represented well in our data. However, there are indications that we have not seen the full diversity of the batch of Intesti we analyzed. A phage cluster amplified on PAO1was barely even present in the sequencing data of the cocktail, confirming that we potentially missed low abundance phages. It is not clear which impact the abundance of a particular phage or phage cluster has on its efficacy in the host, since specific amplification upon encountering the host is an important factor in therapeutic applications.

It is the authors' understanding that the library preparation we used favors dsDNA and the vast majority of phages known today are indeed tailed dsDNA phages [4]. Nevertheless, we cannot exclude the possibility that the cocktail contained ssDNA phages, especially since we introduced a 5 kb size cutoff for contig groups. It is the authors' experience, that contig groups smaller than that may not be true clusters but rather shared modules. At a size smaller than 5 kb it is further difficult to obtain an unambiguous attribution to a certain phage species or cluster of species due to the aforementioned shared modules.

Intriguingly, the three clusters that contain the most common phages in the cocktail, namely D6, D12 and the presumed *Proteus* phage cluster, are also those we know the least about, as they are the ones most different from previously studied phages. For the presumed *Proteus* phage, it is not even sure whether the two contigs form a single cluster, though each by itself is also very abundant (compare depth of coverage and its standard deviation for the *Proteus* phage, Figure 2). We have predicted the phages to belong to the *Siphoviridae* based on tail fibers, but it is not known what their hosts are.

There is a possibility that some of the phage components in the cocktail derive from induction of prophages in the propagating strains, which may explain the comparatively high prevalence of *Siphoviridae* in the Intesti cocktail as well as the presence of lysogeny-related genes. This hypothesis could not be tested since the propagating strains are proprietary and therefore not available.

4.2. Concerning the Synonymous Clusters and Amplification by Bacterial Hosts

It should be remarked that while the clusters infecting each host could be identified, it is not possible to say whether or not all phages in a given cluster are causing infection. In the case of cluster F1, of which only about half were amplified, the distinction was clear.

As was the case in the unamplified cocktail, the depth of coverage varied between contigs belonging to the same phage cluster in the host-amplified samples. This could signify a bias for

amplification of only certain parts of the cluster. On the other hand, chimeric *vs.* non-chimeric contigs can also cause a variation in depth within a cluster (see Section 3.2.2).

Further, it turned out that the phage cluster we presumed to be *Proteus* specific, because of its presence in the Proteus amplified sample and the fact that it did not have any hits to the nr nucleotide database, did not actually cause infection in the *Proteus vulgaris* used in this study. It is therefore unclear what kind of phage those two contigs represent and whether they should be clustered or separate. The only evidence we have is that both of them have high depth of coverage values, which are very similar to each other.

4.3. Comparison to Other Phage Cocktail Studies Employing Metagenomics

McCallin *et al.* published a metagenomic analysis of a Russian phage cocktail intended for treatment of *Escherichia coli*/*Proteus* infections in 2013. Their methodology was somewhat different and more extensive on the experimental side. Our study had its focus on bioinformatics and specifically sequence analysis tools. These kinds of analyses are cheap and fast compared to traditional lab techniques which is why we wished to test their suitability for phage cocktail analysis. Naturally, they do not replace experimental evidence, however we think that by sequencing first and employing bioinformatics prior to further lab work, we are able to gain insight and can design lab experiments more efficiently. This will save time and money, especially as more tools are being developed and databases grow more extensive.

In concordance with the results of McCallin *et al.*, we also observed a great complexity within the cocktail we analyzed. McCallin *et al.* found primarily *Myoviridae* (34%) and *Podoviridae* (24%) in their cocktail. In comparison to that, the Intesti cocktail is also mainly composed of *Myoviridae* (35%), but the second most abundant family was *Siphoviridae*, which were almost as abundant (32%).The cocktail analyzed by McCallin *et al.* is, however, of very different scope, targeting solely *E. coli* and *Proteus*, while the Intesti cocktail we analyzed targets a more broad spectrum of enteric bacteria.

In the *Escherichia coli*/*Proteus* targeting cocktail, McCallin *et al.* identified phages of the *Myoviridae* subfamily *Tevenvirinae* and the genus *Felixounalikevirus*, plus phages of the proposed genus of rv5-like virus, as well as the *Podoviridae* genera *T7likevirus*, *SP6likevirus* and *N4likevirus*. The Intesti cocktail also contained clusters related to those two *Myoviridae* genera and subfamily and the *Podoviridae* genera *T7likevirus* and *SP6likevirus*. The Intesti cocktail appears to have a greater diversity of component phages compared to the Russian cocktail, which is in accord with its broader spectrum of application. As the sequencing data produced in the study of McCallin *et al.* is not publically available, the authors were unable to directly compare the phage clusters identified in the Intesti cocktail to the phages identified in the Russian cocktail.

Neither study identified undesirable genes within the cocktail, but this is not a guarantee for safety since the databases are not exhaustive. The two genes showing homology to integrases warrant further investigation.

When McCallin *et al.* classified their redundancy removed reads with MEGAN, they observed 23% of reads without hits. In comparison, 25% of the redundancy reduced reads in our sample mapped to contigs that could not be assigned, *i.e.*, had no significant BLAST hits. However, McCallin *et al.* compared their reads to the non-redundant protein collection and employed blastx, which has a higher sensitivity. Therefore, the numbers cannot be directly compared between the two studies. Furthermore, when looking at assembled contigs the total size of the contigs which had no database hits, including the putative Proteus phage, was only 16% of the total assembly size, though many of the clusters with known relatives appeared to have novel parts, as evidenced by the fact that their coverage by their database references is not complete (compare Table 2).

Lastly, the metagenomics approach differed between our study and that of the Russian phage cocktail in that we focused on assembling first and subsequently characterizing the contigs we had obtained, while McCallin *et al.* did more characterization work on the read data and with mapping. The main reason we chose direct *de novo* assembly of the full sample is that we were concerned about

creating an artificial separation of the data by relying on mapping, especially since at least some phages are known to be modular and to frequently switch modules, as illustrated for *Staphylococcus* phages by Deghorain *et al.* [43]. Essentially, the focus of our study was on discovery.

4.4. Future Perspectives

One of the purposes of this study was to explore which types of sequence-based analysis are suitable for phage cocktails and whether their results are useful. We hope to ignite discussion on how the analysis of complex phage products can be done in the future.

5. Conclusions

The aim of this study was to identify and analyze the major components of the Intesti phage cocktail. Returning to the question posed in the title, we conclude that a great amount of information can be gained from examining a phage cocktail directly by metagenomic analysis, by relying on databases and bioinformatics tools, though careful interpretation is crucial and not always straight forward. Furthermore, we show that the kind of information presented in this article can be gained without the need to separate and amplify individual phages prior to sequencing, which may not always be possible especially when propagating strains are unavailable or unknown. As databases grow more extensive with sequencing projects on the rise and more tools get developed, we expect that the kind of bioinformatics analysis we employed in this study will grow more powerful and accurate.

Acknowledgments: This work was supported by the Center for Genomic Epidemiology at the Technical University of Denmark and funded by grant 09-067103/DSF from the Danish Council for Strategic Research. The authors would like to thank Finn Kvist Vogensenof the University of Copenhagen for insightful discussions. Furthermore, we would like to express our gratitude towards Nikoloz Nikolaishvili for providing us with the sample of the Intesti phage cocktail.

Author Contributions: Mette V. Larsen conceived, designed and coordinated the project. Katrine G. Joensen and Henrike Zschach performed the laboratory work. Katrine G. Joensen provided guidance for the laboratory work. Barbara Lind hard performed the read mapping and depth of coverage analysis. Henrike Zschach performed the rest of the data analysis. Henrik Hasman supplied bacterial strains and helped to interpret experimental results. Ole Lund provided helpful comments from a more technical programming perspective. Marina Goderdzishvili, Zemphira Alavidze, Irina Chkonia, Guliko Jgenti, and Nino Kvatadzeare the main people responsible for developing and producing this current Intestiphage cocktail. They also aided in the historical description of the development and applications of the Intestiphage cocktail. Elizabeth Kutter aided extensively in the analysis and interpretation of the results and in the editing of the paper. Henrike Zschach wrote the paper draft. Mette V. Larsen performed major editing on the paper and was the main senior supervisor.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

References

1. World Health Organization. *Antimicrobial Resistance Global Report on Surveillance 2014*; World Health Organization: Geneva, Switzerland, 2014.
2. Lu, T.K.; Koeris, M.S. The next generation of bacteriophage therapy. *Curr. Opin. Microbiol.* **2011**, *14*, 524–531. [CrossRef] [PubMed]
3. Chan, B.K.; Abedon, S.T.; Loc-Carrillo, C. Phage cocktails and the future of phage therapy. *Future Microbiol.* **2013**, *8*, 769–783. [CrossRef] [PubMed]
4. Hatfull, G.F. Bacteriophage genomics. *Curr. Opin. Microbiol.* **2008**, *11*, 447–453. [CrossRef] [PubMed]
5. Reardon, S. Phage therapy gets revitalized. *Nature* **2014**, *510*, 15–16. [CrossRef] [PubMed]
6. Harper, D.R.; Anderson, J.; Enright, M.C. Phage therapy: Delivering on the promise. *Ther. Deliv.* **2011**, *2*, 935–947. [CrossRef] [PubMed]
7. Kutateladze, M.; Adamia, R. Bacteriophages as potential new therapeutics to replace or supplement antibiotics. *Trends Biotechnol.* **2010**, *28*, 591–595. [CrossRef] [PubMed]

8. Kutter, E.; Borysowski, J.; Międzybrodzki, R.; Górski, A.; Weber-Dąbrowska, B.; Kutateladze, M. Clinical phage therapy. In *Phage Therapy: Current Research and Applications*, 1st ed.; Borysowski, J., Międzybrodzki, R., Górski, A., Eds.; Caister Academic Press: Poole, UK, 2014; pp. 253–284.
9. Kutter, E.; de Vos, D.; Gvasalia, G.; Alavidze, Z.; Gogokhia, L.; Kuhl, S.; Abedon, S.T. Phage therapy in clinical practice: Treatment of human infections. *Curr. Pharm. Biotechnol.* **2010**, *11*, 69–86. [CrossRef] [PubMed]
10. Abedon, S.T.; Kuhl, S.J.; Blasdel, B.G.; Kutter, E.M. Phage treatment of human infections. *Bacteriophage* **2011**, *1*, 66–85. [CrossRef] [PubMed]
11. Sulakvelidze, A.; Alavidze, Z.; Morris, J.G. Bacteriophage therapy. *Antimicrob. Agents Chemother.* **2011**, *45*, 649–659. [CrossRef] [PubMed]
12. Bruttin, A.; Brüssow, H. Human volunteers receiving *Escherichia coli* phage T4 orally: A safety test of phage therapy. *Antimicrob. Agents Chemother.* **2005**, *49*, 2874–2878. [CrossRef] [PubMed]
13. Sarker, S.A.; McCallin, S.; Barretto, C.; Berger, B.; Pittet, A.-C.; Sultana, S.; Krause, L.; Huq, S.; Bibiloni, R.; Bruttin, A.; *et al*. Oral T4-like phage cocktail application to healthy adult volunteers from Bangladesh. *Virology* **2012**, *434*, 222–232. [CrossRef] [PubMed]
14. Wright, A.; Hawkins, C.H.; Anggård, E.E.; Harper, D.R. A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant *Pseudomonas aeruginosa*; a preliminary report of efficacy. *Clin. Otolaryngol.* **2009**, *34*, 349–357. [CrossRef] [PubMed]
15. Rhoads, D.D.; Wolcott, R.D.; Kuskowski, M.A.; Wolcott, B.M.; Ward, L.S.; Sulakvelidze, A. Bacteriophage therapy of venous leg ulcers in humans: Results of a phase I safety trial. *J. Wound Care* **2009**, *18*, 237–243. [CrossRef] [PubMed]
16. McCallin, S.; Sarker, S.A.; Barretto, C.; Sultana, S.; Berger, B.; Huq, S.; Krause, L.; Bibiloni, R.; Schmitt, B.; Reuteler, G.; *et al*. Safety analysis of a Russian phage cocktail: From metagenomic analysis to oral application in healthy human subjects. *Virology* **2013**, *443*, 187–196. [CrossRef] [PubMed]
17. Brüssow, H. What is needed for phage therapy to become a reality in Western medicine? *Virology* **2012**, *434*, 138–142. [CrossRef] [PubMed]
18. Breitbart, M.; Salamon, P.; Andresen, B.; Mahaffy, J.M.; Segall, A.M.; Mead, D.; Azam, F.; Rohwer, F. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 14250–14255. [CrossRef] [PubMed]
19. D'Herelle, F. *The Bacteriophage*; Williams & Wilkins Company: Baltimore, MD, USA, 1922.
20. Mikeladze, C.; Nemsadze, E.; Alexidze, N.; Assanichvili, T. Sur le traitement de la fievre typhoide et des colites aigues par le bacteriophage de d'Herelle. *La Médecine* **1936**, *17*, 33–38. (In French).
21. Chanishvili, N. *A Literature Review of the Practical Application of Bacteriophage Research*; Nova Science Publishers: New York, NY, USA, 2012.
22. Kuhl, S.J.; Mazure, H. d'Hérelle. Preparation of Therapeutic Bacteriophages, Appendix 1 from: Le Phénomène de la Guérison dans les maladies infectieuses: Masson et Cie, 1938, Paris—OCLC 5784382. *Bacteriophage* **2011**, *1*, 55–65. [CrossRef]
23. Andrews, S. FastQC—A Quality Control Tool for High Throughput Sequence Data. Available online: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/ (accessed on 1 February 2015).
24. Schmieder, R.; Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **2011**, *27*, 863–864. [CrossRef] [PubMed]
25. Laserson, J.; Jojic, V.; Koller, D. Genovo: *De novo* assembly for metagenomes. *J. Comput. Biol.* **2011**, *18*, 429–443. [CrossRef] [PubMed]
26. Zerbino, D.R.; Birney, E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829. [CrossRef] [PubMed]
27. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
28. Deurenberg, R.H.; Stobberingh, E.E. The evolution of *Staphylococcus aureus*. *Infect. Genet. Evol.* **2008**, *8*, 747–763. [CrossRef] [PubMed]
29. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]

30. Aziz, R.K.; Bartels, D.; Best, A.A.; DeJongh, M.; Disz, T.; Edwards, R.A.; Formsma, K.; Gerdes, S.; Glass, E.M.; Kubal, M.; *et al*. The RAST Server: Rapid annotations using subsystems technology. *BMC Genom.* **2008**, *9*. [CrossRef] [PubMed]
31. Besemer, J. GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **2001**, *29*, 2607–2618. [CrossRef] [PubMed]
32. Lobocka, M.; Hejnowicz, M.S.; Gagala, U.; Weber-Dąbrowska, B.; Wegrzyn, G.; Dadlez, M. The first step to bacteriophage therapy: How to choose the correct phage. In *Phage Therapy: Current Research and Applications*; Borysowski, J., Międzybrodzki, R., Gorski, A., Eds.; Caister Academic Press: Norfolk, UK, 2014; pp. 23–67.
33. Samson, J.E.; Magadán, A.H.; Sabri, M.; Moineau, S. Revenge of the phages: Defeating bacterial defences. *Nat. Rev. Microbiol.* **2013**, *11*, 675–687. [CrossRef] [PubMed]
34. Zankari, E.; Hasman, H.; Cosentino, S.; Vestergaard, M.; Rasmussen, S.; Lund, O.; Aarestrup, F.M.; Larsen, M.V. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **2012**, *67*, 2640–2644. [CrossRef] [PubMed]
35. Joensen, K.G.; Scheutz, F.; Lund, O.; Hasman, H.; Kaas, R.S.; Nielsen, E.M.; Aarestrup, F.M. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* **2014**, *52*, 1501–1510. [CrossRef] [PubMed]
36. Santos, S.B.; Kropinski, A.M.; Ceyssens, P.-J.; Ackermann, H.-W.; Villegas, A.; Lavigne, R.; Krylov, V.N.; Carvalho, C.M.; Ferreira, E.C.; Azeredo, J. Genomic and proteomic characterization of the broad-host-range Salmonella phage PVP-SE1: Creation of a new phage genus. *J. Virol.* **2011**, *85*, 11265–11273. [CrossRef] [PubMed]
37. Schwarzer, D.; Buettner, F.F.R.; Browning, C.; Nazarov, S.; Rabsch, W.; Bethe, A.; Oberbeck, A.; Bowman, V.D.; Stummeyer, K.; Mühlenhoff, M.; *et al*. A multivalent adsorption apparatus explains the broad host range of phage phi92: A comprehensive genomic and structural analysis. *J. Virol.* **2012**, *86*, 10384–10398. [CrossRef] [PubMed]
38. Zuo, G.; Xu, Z.; Hao, B. *Shigella* strains are not clones of *Escherichia coli* but sister species in the genus *Escherichia*. *Genom. Proteom. Bioinform.* **2013**, *11*, 61–65. [CrossRef] [PubMed]
39. Sheldon, A.T. Antiseptic "resistance": Real or perceived threat? *Clin. Infect. Dis.* **2005**, *40*, 1650–1656. [CrossRef] [PubMed]
40. Piechowski, M.M.; Susman, M. Acridine-resistance in phage T4D. *Genetics* **1967**, *56*, 133–148. [PubMed]
41. Kawai, M.; Yamada, S.; Ishidoshiro, A.; Oyamada, Y.; Ito, H.; Yamagishi, J.-I. Cell-wall thickness: Possible mechanism of acriflavine resistance in meticillin-resistant *Staphylococcus aureus*. *J. Med. Microbiol.* **2009**, *58*, 331–336. [CrossRef] [PubMed]
42. Wang, F.J.; Ripley, L.S. The spectrum of acridine resistant mutants of bacteriophage T4 reveals cryptic effects of the tsL141 DNA polymerase allele on spontaneous mutagenesis. *Genetics* **1998**, *148*, 1655–1665. [PubMed]
43. Deghorain, M.; van Melderen, L. The Staphylococci phages family: An overview. *Viruses* **2012**, *4*, 3316–3335. [CrossRef] [PubMed]

8 Phage communities in sewage

Another aspect of phage therapy is very relevant yet often taken for granted - where does one find good candidate phages? Traditionally, therapeutic phage have often been isolated from human sewage, which has even been described as the ideal isolation source by Lobocka *et al* [57].

In this study, we took the chance to examine a set of sewage samples that was originally collected to track the spread of antimicrobial resistance genes in different populations around the world. We wanted to use them to look at phage communities in sewage instead. Luckily, the samples had been sequenced quite deeply as metagenomes which encouraged us to try and extract phage sequences from them. Our findings showed those phages to be both extremely diverse and contain a large amount of novel sequence, making for exciting prospects in further studies.

This was also a good chance to test and in the process improve on MetaPhinder, a tool used to identify phage contigs in metagenomic assemblies. MetaPhinder was originally developed by Vanessa Jurtz during her Master's thesis and is now available in new version with extended output.

This study is still ongoing at the time of writing but I chose to include the results so far since I believe it is a vital part of my PhD and thematically ties into the other two papers.

Article

Phage communities in sewage – A metagenomics cross-country perspective

Henrike Zschach [1], Vanessa Jurtz [1], Mette V. Larsen [2], Ksenia Arkhipova [3], Bas Dutilh [3], Rene Hendriksen [4], Frank M. Aarestrup [4], Ole Lund [1], Morten Nielsen [1]

[1] Department of Bio and Health Informatics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; henrike@bioinformatics.dtu.dk (H.Z.), vanessa@bioinformatics.dtu.dk (VJ), lund@bioinformatics.dtu.dk (OL), mniel@bioinformatics.dtu.dk (M.N.)
[2] GoSeqIt ApS, Ved Klaedebo 9, 2970 Hoersholm, Denmark; MVL@goseqit.com
[3] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud university medical centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands; bedutilh@gmail.com (BD), arkhipova.a.ksenia@gmail.com (KA)
[4] Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kemitorvet, 2800, Kgs. Lyngby, Denmark; fmaa@food.dtu.dk (FMA), rshe@food.dtu.dk (RH)

Abstract:

Sewage, a highly competitive and diverse environment, is the primary isolation source for therapeutic phages. However, not much is known about sewage phage communities in different parts of the world. To address this, we have analyzed and compared the phage sequences found in 81 sewage samples from 63 different countries. We also show that MetaPhinder-2.0 is a useful tool for identifying phage sequences in complex metagenomes and is not limited to finding homologs of known phages. Nearly all the phage communities contained a plethora of novel phage sequences independent of their geographic origin, underlining the undiscovered diversity of phages in sewage environment. However, crAssphage was almost universally present. By combining BLASTn hits to full contigs and a tBLASTn search against custom databases of conserved structural phage genes, we were able to assign taxonomic labels on family level to on average 25% of phage reads. We did not observe a clustering of samples by geographic region when comparing their genomic distances as measured by Mash. All samples were highly variable from each other. Further, when investigating the occurrence and coverage of known phages in sewage, we discovered intriguing patterns that corresponded to distinct phage families.

Keywords: phage metagenomics; phage taxonomy; identification of phage contigs in complex environmental samples

1. Introduction

A large proportion of phages intended for therapy are isolated from sewage water which contains many of the major human pathogens and is considered an optimal isolation source [1]. In 2015, Mattila *et al* published a feasibility study on this topic, finding that isolation of phages from sewage was successful identifying phages against Pseudomonas aeruginosa, Salmonella, extended spectrum beta-lactamase Escherichia coli, and Klebsiella pneumoniae. However, it remained difficult to isolate phages against vancomycin resistant Enterococcus and Acinetobacter baumannii as well as methicillin resistant Staphylococcus aureus [2].

Sewage is furthermore a highly competitive environment and a source of untapped biodiversity. For those reasons, there is a great need to learn more about sewage phage communities. However, due to their enormous sequence diversity and absence of common marker genes, phages are not readily identified in mixed metagenomic samples. This is especially true in samples that have

not been specifically treated to amplify viral DNA and remove bacterial and eukaryotic DNA. We here use an updated version of MetaPhinder [3], a tool that identifies phage contigs based on their cumulative average nucleotide identity (ANI) to a database on known phage genomes.

In this study, we have investigated the phage components of sewage samples from around the world. We aim to address issues related to how similar samples are to each other, which proportion of phage contigs we are able to assign taxonomically and to what degree they display similarity to known phages. We further describe an update to MetaPhinder, a tool to identify phage contigs in metagenomic assemblies.

2. Materials and Methods

2.1 Sewage samples

The Global Sewage Surveillance Project has the goal to surveil infectious diseases and antimicrobial resistance in human sewage around the world in order to determine the occurrence and burden of resistance in defined healthy human populations. To that end, the project coordinators have invited countries to collect two liters of urban sewage and send them to the National Food Institute at the Technical University of Denmark (DTU). For more information see http://www.compare-europe.eu/library/global-sewage-surveillance-project. The project is associated with COMPARE (http://www.compare-europe.eu/) and funded by the World Health Organization.

In 2016 the Global Sewage Surveillance Project has collected a total of 81 samples of sewage from 63 different countries. For this study, we received the trimmed reads and full assemblies of those samples. We then identified and extracted the phage contigs in the assemblies for further analysis.

2.2 Metadata

The following metadata was available to us:

Sample location (city, county, and GPS coordinates), sample region, sample site and sample date. We have only made use of the geographic metadata. Figure 1 shows a map of sample locations.

Figure 1. Sampling locations. For two samples only the country was available (marked with a grey star). El Paso, Texas was sampled four times (marked with a blue cross). The map was generated from GPS data using the webtool HamsterMap at http://www.hamstermap.com/custommap.html .

2.3 Reference phage database

We created a database of all available complete phage genomes to use as references. This was done by querying NCBI nucleotide with the search term '(phage [Title]) AND complete genome' as well as downloading the entire database of phage genomes available on phantom.org Genomes from both sources were combined and homology reduced on 100% sequence identity to remove duplicates. The final database comprised 5477 genomes. All downloads were performed on 06. June 2017.

2.4 Sample preparation

The samples were spun down and DNA was isolated using the DNA isolation QIAamp Fast DNA Stool protocol. Subsequently, the samples were sent to Oklahoma Medical Research Foundation for sequencing. Here, DNA was sheared to ~300 bp and the NEXTflex PCR-free DNA-seq library preparation kit was used for library preparation. The samples were multiplexed and sequenced on a HiSeq3000 using 2x150 bp paired end sequencing. Several of the samples were sequenced multiple times.

2.5 Sequencing and assembly

Raw sequencing data was quality trimmed and assembled with SPAdes 3.9.0 [4] using the -meta flag. For the samples that were re-sequenced several times reads from all sequencing rounds were used.

2.6 Update to MetaPhinder, identification of phage sequences and phiX removal

An updated version of MetaPhinder [3], from here on referred to as MetaPhinder-2.0, was used to identify phage contigs within the assemblies. The first version of MetaPhinder was based entirely on the cumulative average nucleotide identity (ANI) of a query contig to a phage database. ANI is hereby defined as:

$$\%ANI = \frac{\sum_{i=1}^{N} percentID_i * al_i}{\sum_{i=1}^{N} al_i} * m_{cov} \quad (1)$$

where $percentID_i$ is the percent identity reported by BLASTn [5], al_i is the alignment length and m_{cov} is the fraction of the query sequence covered by alignment to the reference. N is the number of hits between query and the sequences in the reference database. The cumulative ANI considers hits from any phage genome in the database so long as the E-value of the hit is less or equal to 0.05.

In the updated version of MetaPhinder, we have removed the fixed classification threshold of 1.7% ANI to the phage database. For the sewage dataset studied here we instead used 10% ANI, see results section. Further, we included a comparison of the query contig to a database of 5000 complete bacterial sequences from NCBI's refSeq. Those bacterial sequences were split up into k-mers of length 16. In order to limit database size, we only retained k-mers with prefix ATG. We then removed all k-mers that occur in phages from this bacterial database. To run a comparison, query contigs are also split into k-mers of length 16 and prefixed with ATG. Since phage k-mers have been removed from the bacterial database, the k-mer query coverage becomes a direct measure of how much the query contig resembles a bacterium.

The further analyses described below were run only on the contigs classified as phage by MetaPhinder-2.0. In addition to that, the phage contigs were compared to the sequencing control phiX174 by BLASTn [5] and contigs with greater than 99% identity were removed from the analysis.

2.7 Fraction of phage DNA

The fraction of phage DNA per sample was calculated by dividing the number of base pairs in phage contigs by the total number of base pairs in the assembly.

2.8 Abundance estimation

Following assembly and identification of phage contigs, trimmed reads were mapped to the contigs in order to estimate their abundance. In many of the following analysis, this abundance is expressed as percentage of phage associated reads mapped to a contig.

2.9 Assigning taxonomic labels to phage contigs

We employed two different strategies to assign taxonomic labels to the phage contigs.

Firstly, contigs inherited labels on species as well as family level from their best hit in the database of reference phage genomes, if the average nucleotide identity (ANI) of contig to reference phage was equal to or higher than 80%. Note that this ANI value is only to the top hit phage, not the cumulative ANI to the full phage database. Some known phages such as crAssphage lack a phage family classification. In this case, a contig with the best hit to such a phage was assigned the family label 'unknown'. This should be distinguished from the label 'None' which was assigned to contigs that did not have a reference with ANI >= 80%. With this approach, we identify matches to the full contig.

Furthermore, labels on family level were predicted based on homology to four gene classes generally assumed to have a high conservation rate because they are essential for the correct functioning of the phage particle. These are capsid, baseplate and tail fiber encoding genes, as well as phage associated DNA polymerase genes. A database was constructed for each of those categories by firstly querying NCBI protein with the search term 'capsid AND phage [Title]' and setting the species filter to 'Viruses'. The 'capsid' was replaced with 'baseplate', 'tail fiber' and 'polymerase' for the three other categories respectively. After that, to limit the computation load, databases were homology reduced by using cdhit [6] with a threshold of 90% homology on the shorter sequence. The phage family associated to each known gene was noted. Subsequently, we ran tBLASTn [5] of the databases against the phage contigs in each sample, retaining hits to known structural or polymerase genes if the percent positives was 50% or higher and the alignment length covered at

least 50% of the known gene. 'Percent positives' here refers to the percent of positively scoring amino acids in the alignment, i.e. amino acids that can be substituted for each other according to the BLOSUM matrix. This measure is more sensitive than 'percent identical' in amino acid space. Following this, a phage family label was assigned to the contig if the family label of each known gene found in the contig was identical.

Lastly, we compared phage family labels obtained from both approaches and assigned consensus labels. This was done in the following manner: If only one of the two approaches yielded a taxonomic label this label was the consensus. If both approaches yielded a label and the labels were identical this label was the consensus. If both approaches yielded a label but the label differed, the consensus was set to 'None'.

2.10 Identification of known phages

We further used the similarity function implemented as part of MetaPhinder (see equation 1) to compute ANI values of each known phage to each sample. Those values can be understood as a measure of how much of the known phage's sequence was covered by the sample's phage contigs, as opposed to how much of a contig could be explained by a known phage.

2.11 Genomic distance estimation

Mash [7] was used to calculate pairwise distances between the phage components of the samples. Mash is based on the MinHash principle which allows the reduction of large sequences to representative sketches and has been used to compare for instance webpages and images. In Mash, metagenomes are first reduced to sketches by splitting them into kmers, oligonucleotide stretches of length *k*. All kmers are then hashed with a hash function *h*. A sketch of size *s* contains the *s* smallest hashes returned by *h*. These genome sketches can then be compared by estimating their Jaccard index. See Ondov *et al* for more details.

Since the amount of phage sequence identified in the samples differed considerably, all samples were randomly down sampled to 1100 kb phage sequence 100 times, and 100 all-against-all Mash distances were calculated from those subsets. We then used the average of the 100 Mash distances to obtain one distance for each pair. Average Mash distances within and between regions were further calculated as group averages with the following formulas.

The average distance within one region is:

$$distance\ within\ region = \frac{\sum_{i,j=1}^{N} distance(i,j)_{i<j}}{\frac{N(N-1)}{2}} \quad (2)$$

where $distance(i,j)_{i<j}$ is the distance between samples i and j with index i being less than index j. This is because the Mash distance is symmetric, meaning that $distance(i,j) = distance(j,i)$. N is the number of samples within the region. The denominator of the fraction is the number of combinations.

The average distance between two regions is:

$$distance\ between\ regions = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} distance(i,j)}{N*M} \quad (3)$$

where N is the number of samples in region 1 and M the number of samples in region 2.

3. Results

3.1 Update of MetaPhinder

During the process of finding phage contigs in the sewage samples we have updated the MetaPhinder method to MetaPhinder-2.0. This was done because we observed suspiciously high fractions of phage DNA when predicting with the original MetaPhinder. That first version operated with a classification threshold of 1.7% ANI to a phage database. This threshold was found by setting up a classification task where complete phage genomes and random length pieces of them were mixed with negative data consisting of bacterial, fungal, protozoa, non-phage virus and human sequences. For details see the publication by Jurtz *et al* [3].

In this updated version, we wanted to insure that the contigs we identify as phage were more similar to phages than to bacteria and therefore included a k-mer based comparison to a bacterial database. All k-mers occurring in the phage database had been removed from the bacterial database to account for integrated prophages.

In a metagenomics setting, it may not be advantageous to select a hard %ANI classification threshold, especially since the amount of known phage sequences is still very inadequate compared to their immense diversity. Instead, we opted to extend the output of MetaPhinder-2.0 to give the user as much phage-related information about a contig as possible. For this reason we have removed the classification column in MetaPhinder-2.0's output and added the following columns: k-mer query coverage to bacteria, bacterial top hit, phage top hit, %ANI of the phage top hit, genome size of phage top hit, taxonomic lineage of phage top hit, taxID of phage top hit, host of phage top hit. The user is encouraged to review the presented information and decide on a classification fitting to their dataset. For the sewage data we found that requiring at least 10% ANI to the phage database and a higher ANI to phages than query coverage to bacteria gave good classification results

The web service can be found at https://cge.cbs.dtu.dk/services/MetaPhinder/.

3.2 Fraction of phage DNA

We investigated the fraction of phage DNA present in each sample by dividing the total base pair count of the full assembly by the base pairs assigned to phage contigs. Doing so, we found that the fraction of phage DNA was between 0.83 and 5.33 percent. No influence of the geographic location on the size of the fraction of phage DNA was observed.

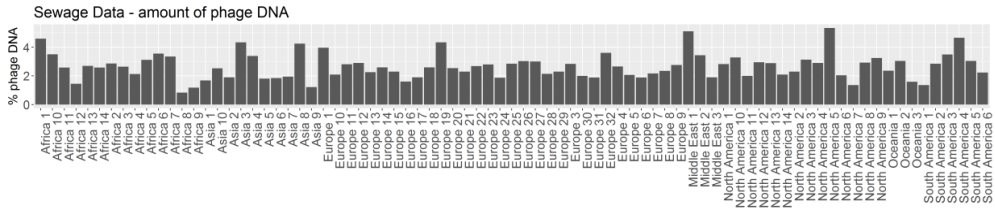


Figure 2. Fraction of phage DNA as percent of total assembled base pairs assigned to phage contigs; displayed per sample and sorted by region.

3.3 Genomic distance estimation

Mash [7] has been shown to be an effective tool for quickly estimating the genomic distances between complex, metagenomic samples based on overlapping kmers. However, a naive application of Mash directly on the recovered phage contigs proved to be heavily biased by the difference in the amount of phage DNA recovered from the different samples. We therefore randomly subsampled each sample to approximately 1100 kb a hundred times, calculated a hundred Mash distances and computed the average distance between samples from that data.

We observed that the majority of samples were equally distant to each other with Mash distances between 0.2 and 0.3, as shown in the resulting heatmap in Figure 3. The four samples taken from El Paso, Texas are encoded as North America 8, 10, 11 and 12. They appear to form a small cluster. However, their pairwise distance is not lower than that of some other samples from distinct locations; compare e.g. Africa 6 and Africa 7.

In addition to that, we have calculated average Mash distances within each region and between regions and observed that there was no substantial difference. This illustrates further that phage communities in samples from the same geographic region are on average not more similar to each other than samples from different regions.

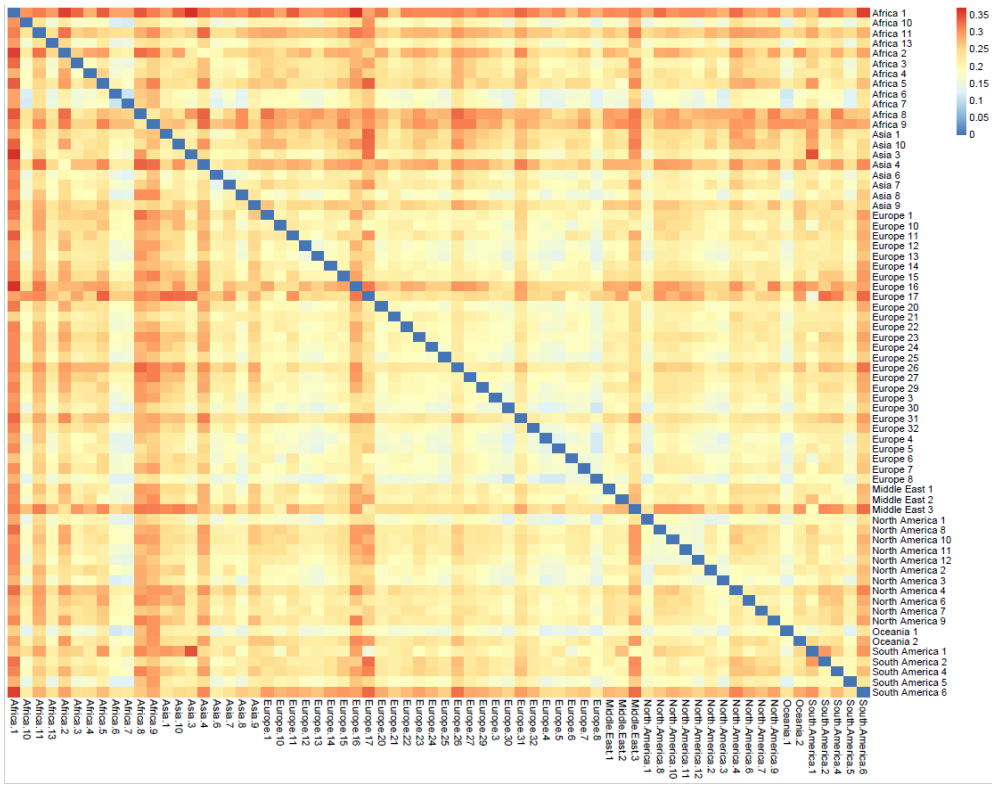


Figure 3. Heatmap of the average Mash distance between all samples. Rows and columns are sorted by region. There appear to be small clusters of higher similarity but overall all the samples are distant to each other.

3.4 Assigning taxonomic labels

3.4.1. Full contig hits

For taxonomic assignment, we first investigated how well the phage contigs in each sample were covered by alignment to known phages. To do so, we divided the contigs into five groups depending on the ANI to their closest reference. The groups were as follows: 0-20% ANI, 20-40% ANI, 40-60% ANI, 60-80% ANI and 80-100% ANI. We then plotted the distribution of those five groups for each sample, see Figure 4. In order to have a better representation of actual abundances, we used the percentage of phage reads mapped to the contigs instead of the percentage of contigs directly.

We found that a large proportion (more than 50% in most samples) of the phage sequences found in sewage were very distant to all known phages, with their best reference yielding ANI values between 0 and 20%. The lowest ANI value identified was only 0.38%. Note, that it is possible for a contig to have a very low ANI to its best reference and still pass MetaPhinder-2.0's classification threshold since MetaPhinder-2.0 accumulates hits across the whole phage database.

We also observed that the proportions between the five groups of contigs did not vary considerably between samples. This means that sewage samples from Europe and North America do not contain an observably higher proportion of known phages than for example African and Asian samples. Notable exceptions to this are the samples Europe 19 and South America 3, in which respectively ~45% and ~30% of reads mapped to contigs with high similarity references. However, upon closer investigation it turned out that both these samples contained a low amount of assembled phage DNA. The phage contigs in Europe 19 amounted to only 137 kb and those in South America 3 to 307 kb sequence.

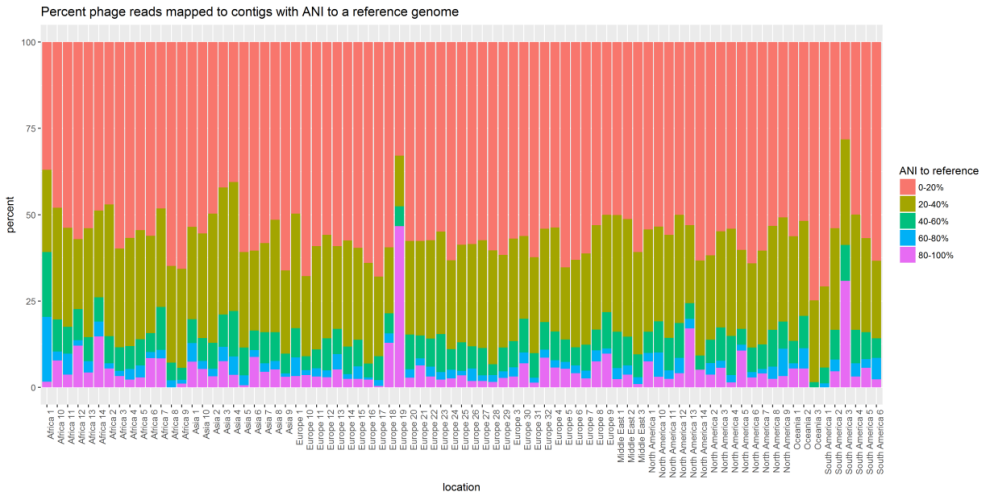


Figure 4. Percent of reads mapped to contigs with 0-20% ANI, 20-40% ANI, 40-60% ANI, 60-80% ANI and 80-100% ANI to their respective best reference phage. The amount of reads with 0-20% ANI varied between 28 and 75% but is generally around 50% or higher.

On the other hand, there was a small proportion of reads mapped to contigs that had very high similarity to their references (ANI values between 80-100%). We extracted these reference phages and investigated whether they were shared across several samples or unique to their sample. The results are displayed in Figure 5. It shows that most reference phages covering a contig with 80% ANI or higher only do so in one to five samples. One reference however is found in 73 out of 81

samples. This was crAssphage. The other phage found to be a good reference in 52 of the samples is Streptococcus phage phiNJ3.

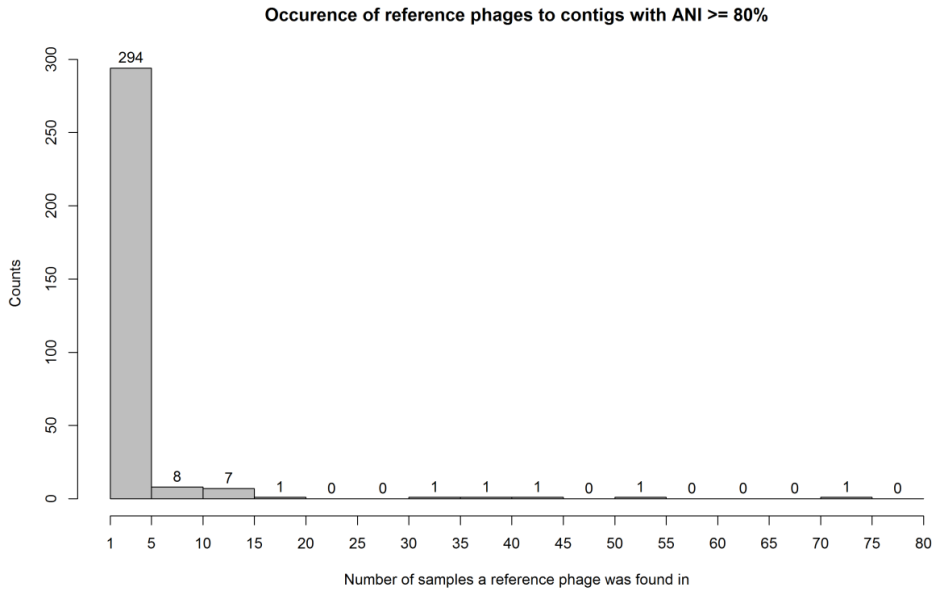


Figure 5. Histogram of the 314 references phages found to cover phage contigs with at least 80% ANI. The height of the bars corresponds to how many samples a reference occurred (with the above-stated ANI threshold). It can be seen that while most references are unique to their sample, two references are found in respectively 52 and 73 samples. They are Streptococcus phage phiNJ3 (in 52 samples) and crAss phage (in 74 samples).

3.4.2 Phage family labels based on conserved genes and consensus labels

In an effort to obtain taxonomic labels for a larger proportion of contigs, we next compared the contigs against three databases of conserved structural phage genes as well as one database of phage-associated polymerase genes. The labels obtained in that way were on family level instead of species level.

This strategy considerably increased the percentage of contigs with a taxonomic label for most samples, see green bars in Figure 6. Once again, we depict the percentage of phage reads mapped to these contigs to better account for abundance. This outcome was expected since the likelihood of finding a single gene match is intuitively greater than the likelihood of finding a match for a whole contig.

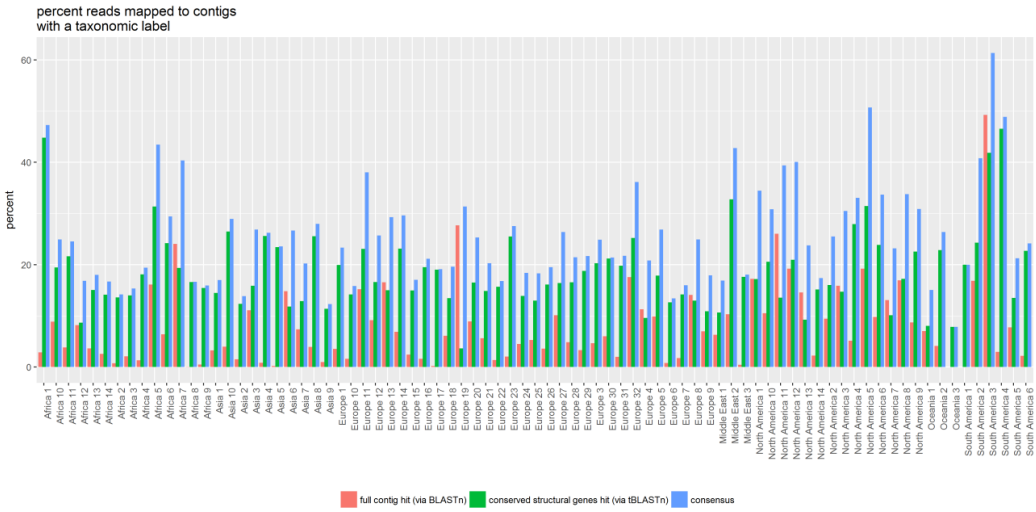


Figure 6. Percent of reads mapped to contigs with a phage family label using two different approaches as well as a consensus. Red: Taxonomic labels inherited from hits to the full contig, extracted from MetaPhinder-2.0 results with ANI >= 80%. Green: Taxonomic labels via matching to database of conserved structural genes. Blue: Consensus.

In some samples however, the percent reads with a label actually decreased. This can for example be observed in sample Asia 6 and Europe 19. The reason for this is contigs mapped to a phage whose structural genes were at the time of writing not available in genbank and therefore not part of our database of conserved structural genes. One of these phages that features prominently in our data is crAssphage. Such contigs can only receive a taxonomic label from their best full contig hit.

From this result, we decided to make a consensus of both approaches as described in the Methods part. Doing so further increased the percent of reads for which we could assign a taxonomic label to the point where we have labels on phage family level for at least 15% of phage reads in most samples and up to more than 40% in a few samples. This increase is largely due to complementary results, i.e. contigs that only obtained a label in one of the predictions but not the other.

3.5 Identification of known phages

In addition to trying to classify the phage contigs, we also sought to find out which of the known phages are present fully in the samples. For this, we once more used the similarity function of MetaPhinder-2.0 but swapped query and database. In that way, we calculated ANI values for each known phage to each sample, thereby describing how well the known phage is covered by the sample. The result is shown in Figure 7.

This figure consists of a boxplot of the ANI values observed in all 81 samples per known phage. Only phages that were covered with an ANI of at least 50% in at least one sample are included. We have also color coded the phage labels on the y-axis by their families: *Siphoviridae* (red), *Myoviridae* (green), *Podoviridae* (blue) or unknown (grey). It can be seen that there appears to be pattern in the distribution of ANI values.

At the top of the plot are crAssphage and Bas gut phage, a variation of crAssphage, which were present in almost every sample with an average ANI of 75%.

373
374 Below, we see a second group of phages that had an ANI between 0 and 10% in the majority of
375 samples. However there are heavy tails to the right of the distribution, as evidenced by the long
376 whiskers of the boxplots. Most of the phages in this group were members of the *Siphoviridae*.
377
378 The third group of phages was not present in the majority of samples but create a curious wave
379 pattern of eight lines at between 3 and 65% ANI. Each of these lines was one sample, which we have
380 verified by coloring the ANI values belonging to the same sample in the same color (only for phage
381 group three). The lines are caused by the fact that all of those phages had nearly the same ANI value
382 to a certain sample. This group was dominated by *Myoviridae* phages.
383
384 Finally, the fourth group of phages was also not present in most samples but if they were, their
385 ANIs were quite high between 50 and 90%. In this group, members of all three phage families were
386 found as well as a few phages of unknown family.
387
388

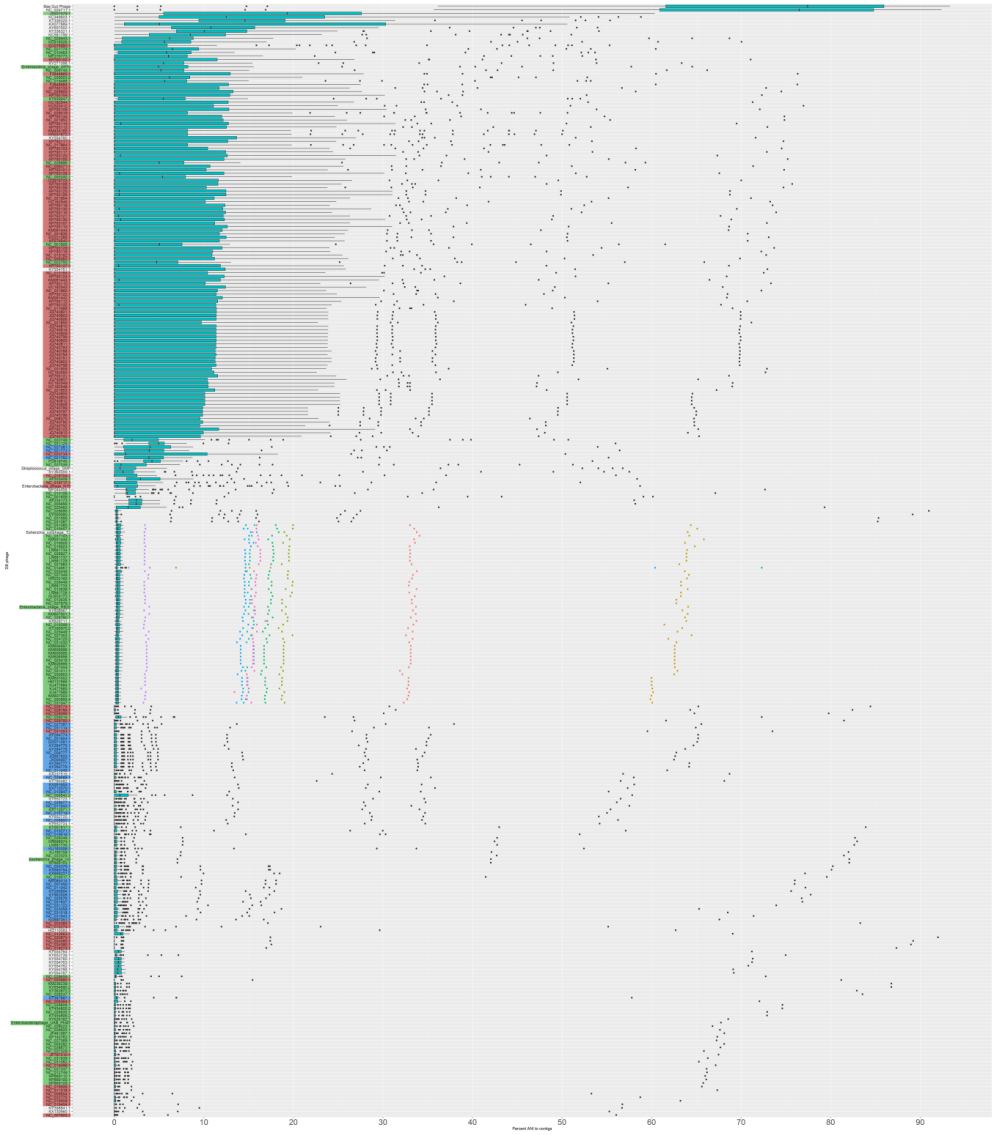


Figure 7. ANI of known phages to sewage samples. Each row corresponds to one known phage and shows a boxplot of that phage's ANI values to each sewage sample, i.e. how well this phage is covered by alignment by the samples' contigs. It can be seen that two versions of crAssphage are present in almost every sample with high ANI of at least 60%. Further, there appear to be three groups of phages that follow similar patterns. Phage labels are colored in accordance to their family: Myoviridae (green), Siphoviridae (red), Podoviridae (blue) or unknown (grey). Only phages with an ANI greater than 50% in at least one sample are shown.

4. Discussion

We have analysed the phage sequences found in metagenomics assemblies of 81 samples of sewage. We found that the phage communities differ considerably between samples and contain

many sequences that are either novel or distant from those of the phages currently available in public databases.

This outcome is not unexpected. Metagenomic studies often find large fractions of sequences that do not map to the current databases, conventionally referred to as 'biological dark matter' [8]. Those sequences are attributed to uncultured bacteria, archaea and viruses, among them bacteriophages. This is in line with the finding that 85-99% of bacteria and archaea can currently not be grown in the lab [9]. Naturally, that means that also their phages cannot be cultivated and will thus be missing from the databases. Perez-Sepulveda *et al* for example reported that the majority of phages in marine environments are part of the sequencing dark matter, i.e. reads that cannot be mapped to known genomes [10].

We further found the fraction of phage DNA was between 0.83 and 5.33 as measured by the percent base pairs assigned to phage contigs. This is congruent with findings from Reyes *et al* who report that viral DNA makes up 2-5% of the total in most environments [11].

In respect to matching the phage contigs in the sewage sample to known phages, we found that only a small percentage of contigs map to a reference phage with a high %ANI and that most of those references are only found in few samples. This again ties in with the notion that the currently known phage sequences hardly even begin to cover the space of phage sequence diversity. CrAssphage is the major exception to that. This phage has already been described as highly abundant in the paper that describes its discovery by Dutilh *et al* [12]. This notion was further confirmed by studies on the human gut metagenome/phageome by Yarygin *et al* [13] and Manrique *et al* [14]. When looking at these results, we further need to consider that some phage contigs were quite short (<= 10 kb). This makes it statistically more likely to obtain an ANI of 80% or higher from a hit to only a short region in the reference phage.

Our results on how well known phages are represented in the sewage samples showed a grouping of phages into four distinct patterns of ANI value distributions. Two of those groups are each dominated by a single phage family, though we are unsure how to interpret this result.

Group three, dominated by *Myoviridae*, showed a wave-like pattern of ANI values far out from the mean of the distribution, which was close to 0. The pattern is probably caused by the phages in that group being closely related to each other. It is conceivable that the same group of contigs align equally well to each of those known phages, causing them all to have a very similar ANI value.

Regarding the MetaPhinder update, we decided to remove a hard %ANI threshold, include more information on the phage top hit and report relatedness to bacteria as well.

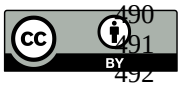
We reason that in an actual metagenome, the original %ANI threshold may be too permissive because of stray hits to genes shared for example between phages and bacteria. The presence of integrated prophages that are not annotated as such in bacterial genomes makes it quite difficult to differentiate between a phage contig and a bacterial contig especially with short contigs. We address this issue by comparing to a bacterial database that had all known phage k-mers removed from it.

Further, while employing two different measures of comparison may seem counterintuitive we consider our approach to have merit. ANI and k-mer query coverage are not directly comparable, however both give an indication of genomic relatedness and have different advantages. The original MetaPhinder paper has shown that for phage classification %ANI is a better measure than k-mer query coverage. At the same time it is computationally very expensive to calculate %ANI of a contig to a large bacterial database, also since bacteria genomes are on average 10x times longer than phage

genomes. Using k-mers furthermore allowed us to efficiently remove phage-like sequences from the bacterial database without having to cut them out of the bacterial genomes. Since we are not interested in finding the best bacterial match to a contig but merely in estimating whether the contig is more similar to bacteria than phage, we argue that our divided approach is applicable and reasonable.

References

1. M. Lobocka, M. S. Hejnowicz, U. Gkagała, B. Weber-Dąbrowska, G. Wkegrzyn, and M. Dadlez, "The first step to bacteriophage therapy – how to choose the correct phage," in *Phage Therapy: Current Research and Applications*, J. Borysowski, R. Miedzybrodzki, and A. Górski, Eds. Norfolk: Caister Academic Press, 2014.
2. S. Mattila, P. Ruotsalainen, and M. Jalasvuori, "On-demand isolation of bacteriophages against drug-resistant bacteria for personalized phage therapy," *Front. Microbiol.*, vol. 6, no. NOV, pp. 1–7, 2015.
3. V. I. Jurtz, J. Villarroel, O. Lund, M. Voldby Larsen, and M. Nielsen, "MetaPhinder - Identifying bacteriophage sequences in metagenomic data sets," *PLoS One*, vol. 11, no. 9, pp. 1–14, 2016.
4. S. Nurk *et al.*, "Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads," Springer, Berlin, Heidelberg, 2013, pp. 158–170.
5. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–10, Oct. 1990.
6. L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.
7. B. D. Ondov *et al.*, "Mash: fast genome and metagenome distance estimation using MinHash.," *Genome Biol.*, vol. 17, no. 1, p. 132, 2016.
8. R. J. Robbins, L. Krishtalka, and J. C. Wooley, "Advances in biodiversity: metagenomics and the unveiling of biological dark matter.," *Stand. Genomic Sci.*, vol. 11, no. 1, p. 69, 2016.
9. C. Lok, "Mining the microbial dark matter," *Nature*, vol. 522, no. 7556, pp. 270–273, Jun. 2015.
10. B. Perez Sepulveda, T. Redgwell, B. Rihtman, F. Pitt, D. J. Scanlan, and A. Millard, "Marine phage genomics: the tip of the iceberg.," *FEMS Microbiol. Lett.*, vol. 363, no. 15, Aug. 2016.
11. A. Reyes, N. P. Semenkovich, K. Whiteson, F. Rohwer, and J. I. Gordon, "Going viral: next-generation sequencing applied to phage populations in the human gut," *Nat. Rev. Microbiol.*, vol. 10, no. 9, pp. 607–617, Aug. 2012.
12. B. E. Dutilh *et al.*, "A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes," *Nat. Commun.*, vol. 5, pp. 1–11, Jul. 2014.
13. K. Yarygin *et al.*, "Abundance profiling of specific gene groups using precomputed gut metagenomes yields novel biological hypotheses," *PLoS One*, vol. 12, no. 4, p. e0176154, Apr. 2017.
14. P. Manrique, B. Bolduc, S. T. Walk, J. van der Oost, W. M. de Vos, and M. J. Young, "Healthy human gut phageome," *Proc. Natl. Acad. Sci.*, vol. 113, no. 37, pp. 10400–10405, Sep. 2016.



9 Host-genomic determinants of Phage susceptibility in S. aureus

Staphylococcus aureus, especially the methicillin-resistant strains, is a growing global health concern and it is not surprising that considerable efforts have been expended into research around phage therapy prospects for this important nosocomial pathogen. The project presented here takes a closer look at one of the principle aspects of phage therapy: What defines whether a given *S. aureus* isolate is sensitive to a given phage?

To address this question we primarily needed three things: A set of clinically relevant *S. aureus* isolates, a set of therapeutic phages and a mathematical model of the interaction. Now this is a bioinformatics department, so we have no lack of models but a distinct lack of biological organisms. It was a good thing both Henrik Westh from Hvidovre Hospital and Ryszard Międzybrodzki from the Hirszfeld Institute were happy to collaborate on this project and so we were able to set up a wet-lab experiment and produce data fitting to our research question.

With this targeted approach, we were able to identify 167 gene families in the accessory genome of *S. aureus* that influence its susceptibility towards the therapeutic phages used by the Hirszfeld Institute. This work is an important step in the direction of well-informed therapy with monovalent phage preparations, especially in a context where DNA sequencing of the causative agent of a severe infection is set to become increasingly common. Methods such as ours can aid in suggesting the appropriate phage tailored to the infecting bacterial strain.

antibiotics

MDPI

Article

Host-genomic determinants of phage susceptibility in MRSA

Henrike Zschach[1*], Mette V. Larsen [2], Henrik Hasman[3], Henrik Westh[4], Morten Nielsen[1*], Ryszard Międzybrodzki[5,6], Ewa Jończyk-Matysiak[5], Beata Weber-Dąbrowska[5] and Andrzej Górski[5,6]

[1] Department of Bio and Health Informatics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark; henrike@bioinformatics.dtu.dk (H.Z.), mniel@bioinformatics.dtu.dk (M.N.)
[2] GoSeqIt ApS, Ved Klaedebo 9, 2970 Hoersholm, Denmark; MVL@goseqit.com
[3] Department of Bacteria, Fungi and Parasites, Statens Serum Institut, 2300 Copenhagen S, Denmark, henh@ssi.dk
[4] MRSA KnowledgeCenter, Department of Clinical Microbiology, Hvidovre Hospital, DK-2650 Hvidovre, Denmark; Henrik.torkil.westh@regionh.dk
[5] Bacteriophage Laboratory, Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, 53-114 Wroclaw, Poland; mbrodzki@iitd.pan.wroc.pl (RM), ewa.jonczyk@iitd.pan.wroc.pl (EJM), weber@iitd.pan.wroc.pl (BWD), agorski@ikp.pl (AG)
[6] Department of Clinical Immunology, Transplantation Institute, Medical University of Warsaw, 02-006 Warsaw, Poland

* Correspondence: Morten Nielsen, mniel@bioinformatics.dtu.dk; Tel.: +45 45 25 24 25

Academic Editor: name
Received: date; Accepted: date; Published: date

Abstract:

Staphylococcus aureus is a major agent of nosocomial infections. Especially in methicillin-resistant strains, conventional treatment options are limited and expensive, which has fueled a growing interest in phage therapy approaches recently.

We have tested the susceptibility of 207 clinical *S. aureus* strains to 12 (nine monovalent) different therapeutic phage preparations and subsequently employ linear regression models to estimate the influence of individual host gene families on resistance to phages. Specifically, we use a two-step regression model setup with a preselection step based on gene family enrichment.

We show that our models are robust and capture the data's underlying signal by comparing their performance to that of models build on randomized data. In doing so, we have identified 167 gene families that govern phage resistance in our strain set and performed functional analysis on them. This revealed genes of possible prophage or mobile genetic element origin, along with genes involved in restriction-modification and transcription regulators, though the majority were genes of unknown function.

This study is a step in the direction of understanding the intricate host-phage relationship in this important pathogen with the outlook to targeted phage therapy applications.

Keywords: phage therapy; bacterial phage resistance; regression modeling; MRSA

1. Introduction

Methicillin-resistant *Staphylococcus aureus* (MRSA) is a growing health concern. It is the agent of many chronic bacterial infections in hospitals as well as in the community. Its resistance to

beta-lactamases severely limits treatment options, drives up the price for therapy, increases unwanted side effects and leads in many cases to worse clinical outcomes [1]. MRSA has been classified as a high priority pathogen on the 2017 list of antibiotic-resistant priority pathogens published by the World Health Organization [2]. Pathogens on this list are considered to pose the greatest threat to human health and to require urgently discovery and development of new antibiotics.

Phage therapy has been proposed as a promising substitute for conventional antibiotics or a co-treatment in the treatment of multi-resistant bacterial pathogens [3]–[7]. Of the *S. aureus* phage known to date, most are temperate phages and belong to the *Siphoviridae* family [8]. Strictly lytic staphylococcal phages, as are typically required for therapy, are almost exclusively found in the *Podoviridae* and *Myoviridae* families [8].

The Hirszfeld Institute of Immunology and Experimental Therapy of the Polish Academy of Science in Wroclaw (HI) has been producing staphylococcal phages for therapeutic purposes since the seventies of the last century [9]. At present its collection consists of nine polyvalent staphylococcal phages (see: Materials and Methods) [10]. Those phages are used at the Phage Therapy Unit in Wrocław under the rules of a therapeutic experiment to conduct treatment of patients with chronic bacterial infections resistant to antibiotic therapy. The result have been encouraging as a good response has been observed in one third of patients [6].

However, in order for phage therapy to be efficient, it is necessary to have a good understanding of the specific interaction between phage and host. There are many strategies by which bacteria aim to evade predation by phages, which is a significant fitness factor and therefore under high evolutionary pressure. Some of the common general phage resistance mechanisms described are: modification of receptor sites to mask them against phage adsorption, restriction-modification systems, abortive infection systems, and CRISPR, to name a few [11].Restriction-modification is a two-part system composed of a methylase and a nuclease. The methylase introduces specific modifications on the organism's DNA, thereby marking it is as self. DNA lacking those modifications, i.e. DNA of foreign origin, will be cleaved by the nuclease. Abortive infection occurs when the host cell recognizes the phage infection before completion of the phage's reproductive cycle and initiates cell death, thereby preventing the phage from successfully creating progeny. CRISPR, an acquired bacterial defense system based on retention and subsequent recognition of fragments of foreign DNA[11], is not typically found in *S. aureus* [12].

S. aureus is known to have a rather large accessory genome that can make up as much as 25% of total genome size [8]. We therefore hypothesize in this study that *S. aureus* may be carrying accessory genes that encode various mechanisms that are geared toward phage resistance. Presence of such mechanisms may hamper the efficacy of phage therapy and it is therefore important to study these in order to perform optimization of phages used for treatment.

Within the phage therapy community, it is being debated whether targeted single phages or cocktails composed of many phages with complementary host ranges are preferable for treatment [13]. Similar to broad-spectrum antibiotics, cocktails can be applied based on the symptoms of the patients, even though the infecting agent has not been isolated or characterized. On the other hand - like broad-spectrum antibiotics - this approach is likely to promote the development of resistance among the bacteria, both the ones causing the disease as well as by-standers. While it is expected that the use of targeted single phages would lead to far less development of resistance, successful treatment is dependent on detailed knowledge of the infecting agent coupled with a thorough understanding of the rules governing the phage-bacteria interaction. With the advent of cheap high-throughput sequencing methods, it is becoming increasingly common to determine the entire genome of infecting bacteria.

In this study, we seek to elucidate the interactions between *S. aureus* and therapeutic phage preparations from the HI with a focus on single phages. To that end, we have tested the susceptibility of a collection of clinical MRSA isolates towards a collection of staphylococcal phage preparations from HI. Both the bacterial and phage collections we used are of great relevance to the phage therapy efforts, since the phages are either already in use or under consideration for experimental therapy in accordance with EU rules concerning compassionate use. Furthermore, the bacterial isolates were obtained from patients showing complicated nosocomial MRSA infections. This strain set represents the most prevalent clonal complexes observed in Denmark and may therefore not be representative of MRSA in different settings.

The genomes of the bacterial strains were determined by whole genome sequencing and through employing a number of bioinformatics tools and machine-learning methods, we attempted to shed light on the genes of MRSA that play a role in determining the susceptibility or resistance towards phages.

2. Results

2.1 General results of the susceptibility testing

A total of 207 MRSA strains were successfully tested for susceptibility to 12 phage preparations. The ratio of susceptible to resistant strains differed between the preparations. The percentage of sensitive strains ranged from 19% to 68% as can be seen in Table 1. We did not observe a large difference in efficacy between single phage preparations and mixtures. However, the efficacies of the different preparations are not directly comparable, since the titer of the phage preparations was not known. Instead the data presented in Table 1 may serve as an indication of whether or not there was sufficient positive and negative data to model the response.

Table 1. Wet lab results of susceptibility testing. All phage preparations were tested at RTD, see Methods. MS-1, OP_MS-1 and OP_MS-1_TOP are mixtures of P4/6409, A5/80 and 676/Z.

Phage preparation	Percent sensitive	Percent resistant
1N/80	31.9%	68.1%
676/F	50.7%	49.3%
676/T	68.1%	31.9%
676/Z	40.6%	59.4%
A3/R	18.8%	81.2%
A5/L	47.3%	52.7%
A5/80	55.1%	44.9%
P4/6409	37.7%	62.3%
phi200/6409	44.0%	56.0%
MS-1	33.8%	66.2%
OP_MS-1	38.6%	61.4%
OP_MS-1 TOP	39.6%	60.4%

2.2 Genetic diversity of the strain collection

Genetic distance between the MRSA strains was measured as 1-orthoANI (see Methods), and the result is depicted in form of a heatmap in Figure 1. This figure reveals a clear clustering of strains into groups with high identity, which follows the established clonal complexes and sequence types of *S. aureus* [14]. Based on this clustering, the strains were split into 5 partitions by visual inspection.

Partition 1 is substantially larger than the other four. This is due to the fact that the strains belonging to clonal complexes CC1, CC5, CC8 and CC80 have a high degree of identity to each other, compare large blue area in the upper left corner. Partitions 2 and 3 are well defined, encompassing CC22 and CC30 respectively. Partition 4 is made up of CC45 and CC398. CC398 is known for its prevalence in swine and cattle. Those strains are genetically distant from the rest of the strains though there is some degree of similarity to CC30. Partition 5 is composed of two clusters of related strains, as indicated in Figure 1. It contains a number of rarer CCs that also show a comparatively high distance in terms of orthoANI to the rest of the data set.

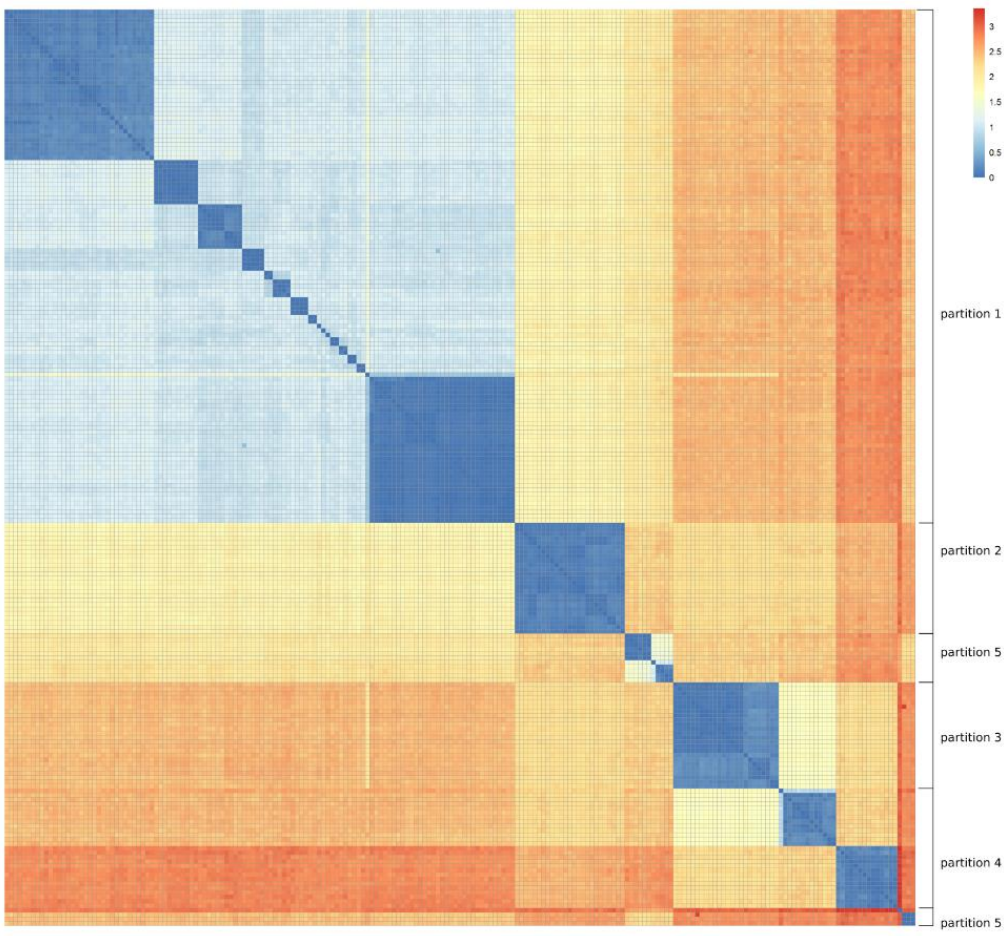


Figure 1. All-against-all matrix of the genetic distance between the 207 MRSA strains used for this study. Distance is calculated as 1 - orthoANI and represented as color, where blue corresponds to lower and red corresponds to greater distance. The assignment of strains to partitions is marked on the right margin.

2.3 Identification of gene families

When predicting and clustering genes, we identified a total of 6419 gene families in the MRSA strain dataset. The distribution of these gene families across the 207 MRSA strains can be seen in

Figure 2, which shows a histogram of abundances of the gene families. 1777 gene families were identified in all 207 strains. These are the housekeeping genes. Furthermore, there is a heavy tail of gene families that were only observed in few strains (left side of the histogram).

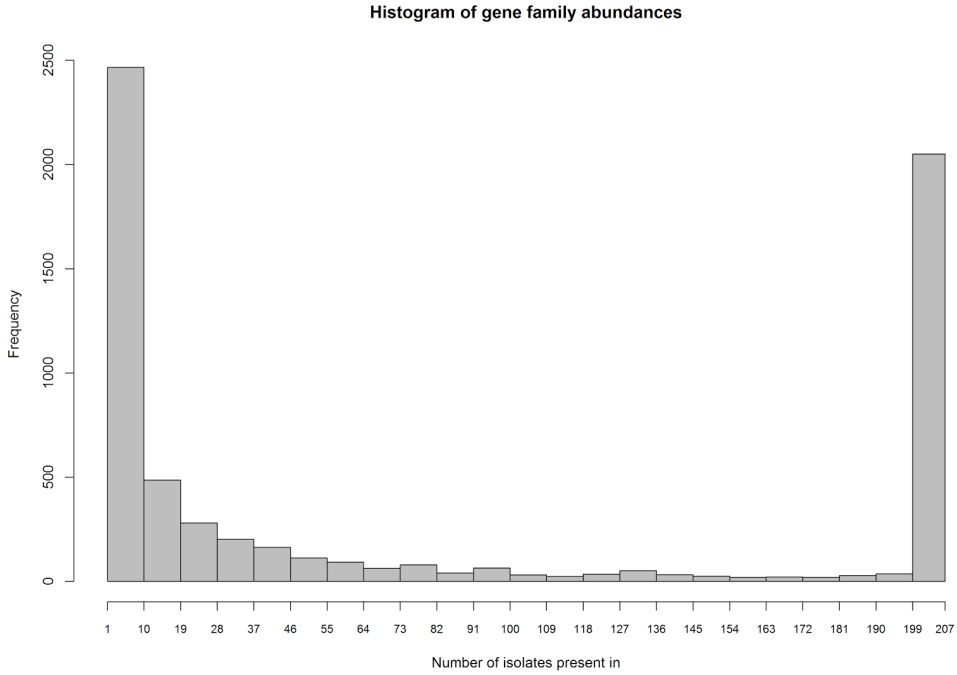


Figure 2. Abundance of gene families in the 207 strains. The peak depicted in the histogram is slightly higher than the number of housekeeping genes, 1.777, since the bin is wider than 1.

2.4 Model construction and feature selection

To identify gene families and construct a model capable of predicting the susceptibility of a MRSA strain to a given phage, a feature selection procedure based on enrichment scores and training of linear regression models was applied. In short, gene families were identified in a two-step procedure, first through a simple enrichment/association test, and second through a refinement step based on regression models combined with consistency constraints.

2.4.1 Enrichment/association test

For each cross validation fold, each gene family was assigned a p-value calculated from its corresponding contingency table estimated once from the original data and once from permuted data. When plotting the distributions of these p-values, illustrated in Figure 3 for the phage P4/6409, we can make several observations:

a) In most phage interactions there is a small tail of gene families with very low p-values, while the majority of gene families have non-significant p-values.

b) In the permuted data, this tail vanishes as was to be expected. We also observed that the p-value distributions of phages 1N/80, A3/R and cocktail MS-1 resemble those of the permuted data much more than those of the real data (see Supplementary Figure S2). This indicates there were not enough positive examples of lysed strains to produce a signal that is distinguishable from random.

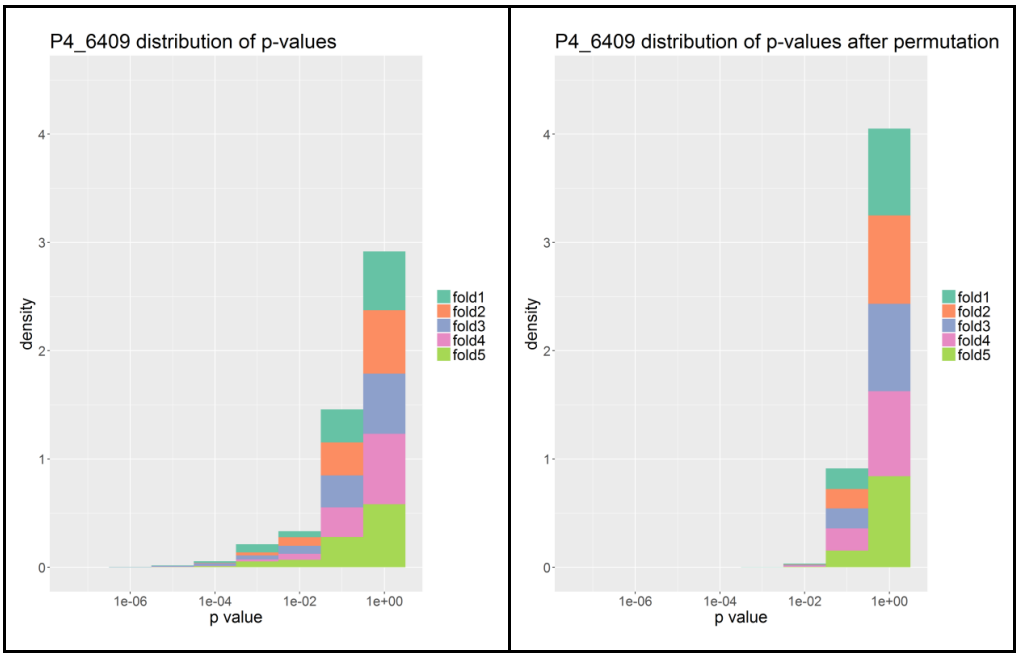


Figure 3. Stacked histogram of p-value distributions across the five folds for the interaction with phage P4/6409. The density is shown instead of counts to account for fold 1 having a 100 times less p values compared to the other folds, since it does not include partition 1 and therefore did not need to be subsampled. Left: Real data. Right: Permuted data.

Based on these observations, a p-value threshold of 0.01 or lower was implemented to admit gene families to the second round of feature selection by regression weights (for details see materials and methods). As seen in Table 2, the number of gene families picked by enrichment varied both by fold as well as by phage. In preparations 1N/80, A3/R and mix MS-1, the number of gene families picked was very low. Further, as expected, we find that no or only very few gene families are selected when analysing the permuted data.

2.4.2 Refinement based on regression models

In the second step of feature selection, we employed linear regression models fitted using Ridge regression. An internal cross validation was used to identify the optimal parameter for the Ridge penalty lambda. The optimal lambda penalty value across the different folds in the cross validation were comparable, indicating that the models are robust, though the size of the feature space varies (see Supplementary Figure S1).

Due to the 5-fold cross validation setup, each gene family was assigned 5 regression weights, which may be NA (not applicable) if the gene family was not chosen by enrichment for that fold. Weights can be either positive or negative. As we chose to model susceptibility as the positive outcome and resistance as the negative outcome, this means that positive weights point towards increased susceptibility, while negative weights point towards increased resistance.

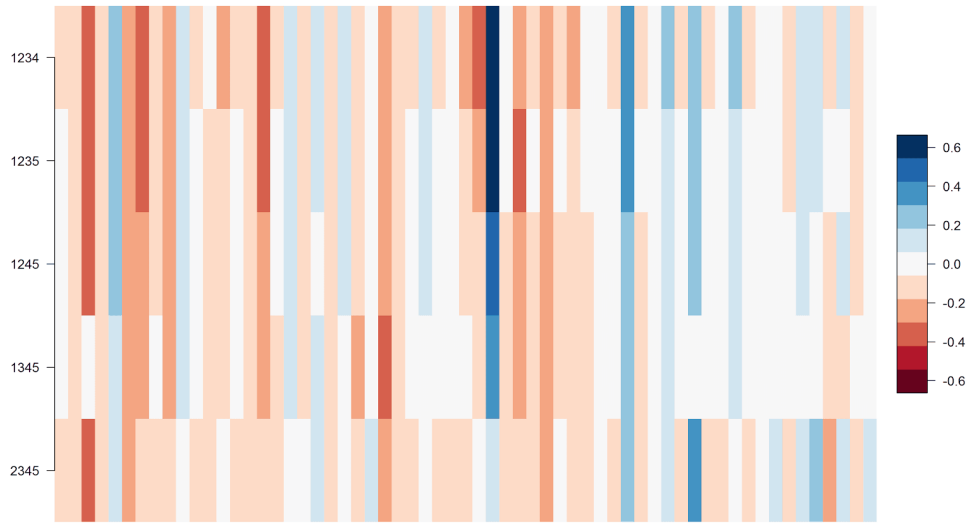
We next required that a gene family should have absolute regression weights greater than 0.01 in at least three of the five partitions to have passed a second selection step. The number of gene families selected in this manner is listed per phage on the right side of Table 2. We term this the set of significant gene families for a certain phage. The number of significant gene families in interaction with phages 1N/80, A3/R and mix MS-1 was too small to train a final model. For the remaining phages, the amount of significant gene families varied between the different phages, though the sets were comparable in size with the smallest comprising 13 and the largest 80 gene families, see Table 2. In total, there were 167 significant gene families. When performing the same procedure on permuted data, significant gene families could only be identified in four phages and a final model could only be trained for two.

Table 2. Summary of the modelling results for real and permuted data. The 'First model' section reports the results of the first filtering procedure based of association analyses. The 'Final model' section gives the result of the second filtering procedure based on regression model fitting combined with consistency constraints. The AUC (area under the curve) is used as performance measure of the final model. The number of gene families selected given in the left part of the table is calculated as the average ± standard deviation across the five folds. If less than two gene families were selected based on regression weights, a final model could not be trained and the associated AUC is reported as NA (not applicable).

	First model		**Final model**			
	Real data	**Permuted data**	**Real data**		**Permuted data**	
Phage Preparation	**No. of gene families selected by enrichment**	**No. of gene families selected by enrichment**	**No. of gene families selected on regression weights**	**AUC**	**No. of gene families selected on regression weights**	**AUC**
1N/80	10 ± 16	0	2	NA	0	NA
676/F	222 ± 144	0	45	0.78	0	NA
676/T	361 ± 243	12 ± 11	79	0.87	3	0.63
676/Z	112 ± 87	11 ± 14	31	0.72	4	0.61
A3/R	13 ± 26	0	1	NA	0	NA
A5/L	184 ± 124	0	37	0.8	0	NA
A5/80	265 ± 148	0	80	0.78	0	NA
P4/6409	200 ± 137	2 ± 4	61	0.79	0	NA
phi200/6409	160 ± 138	0	56	0.79	0	NA
MS-1	6 ± 10	0	0	NA	0	NA
OP_MS-1	86 ± 78	0	29	0.65	0	NA
OP_MS-1_TOP	54 ± 52	1 ± 1	13	0.67	0	NA

222 2.4.3 Final model

223 Final models were next retrained including only the significant gene families passing both
224 selection criteria as input features. Plots of the regression weights assigned by those final models
225 showed the direction of weights to be consistent across folds, i.e. gene families are either found
226 consistently to have positive or negative weights across most of the 5 partitions. This is depicted for
227 the example of phage P4/6409 in Figure 4.
228



229

230 **Figure 4.** Heat map of the regression weights for the final model of phage P4/6409. Columns are gene
231 families, rows are cross validation folds. The color indicates the value and direction of each weight,
232 with blue being strongly positive and red being strongly negative. Weights with low values are
233 white. Results were comparable for other phages with the exception of 1N/80, A3/R and mix MS-1
234 (see Table 2).

235 Out of all the 167 gene families, in total 99 increased phage resistance, 63 increased phage
236 susceptibility and five were ambiguous, meaning that they increased resistance to some phages but
237 susceptibility to others. This confirms that the vast majority of significant gene families identified
238 were consistent in their direction of influence.
239
240 The definition of phage susceptibility we used in this analysis encompasses only the two
241 highest lysis levels, namely confluent lysis and semi confluent lysis. We have re-run the modeling
242 process including also the weakly sensitivity levels and found no difference in the modeling results.
243 This is probably because intermediate sensitivity was rarely observed in our strain set (see
244 Supplementary Table S4).
245

246 *2. 5 Functional annotation of the significant genes*

247 We further sought to characterize the function of the identified significant gene families by
248 comparing them to the eggNOG database. The distribution of functional annotation terms identified
249 for the full set of significant genes is shown in Figure 5, and shows that it was possible to identify a

match in eggNOG for only 60% of gene families. Most genes had either no hit in the eggNOG database or a hit to a NOG of unknown function.

Case-by-case inspection of the functional annotation terms retrieved from both RAST and eggNOG for the 167 significant gene families identified 13 gene families that have terms directly related to phages, while another 18 were related either to other mobile genetic elements such as genomic islands and transposons or to processes associated to them such as transposase activity. Four additional gene families appeared to be part of restriction-modification systems and six had hits to transcriptional regulators.

Out of these groups, only the gene families related to restriction-modification systems were found to consistently be associated with resistance to phage infection (as measured by the sign of the weights in the final model described earlier). The others groups encompass both gene families promoting resistance and families promoting susceptibility, further pointing to the complexity of the host-phage interaction. The full list of annotation terms for all significant gene families can be found in the Supplementary Table S2, together with the gene family's average regression weight across the five cross validation folds per phage.

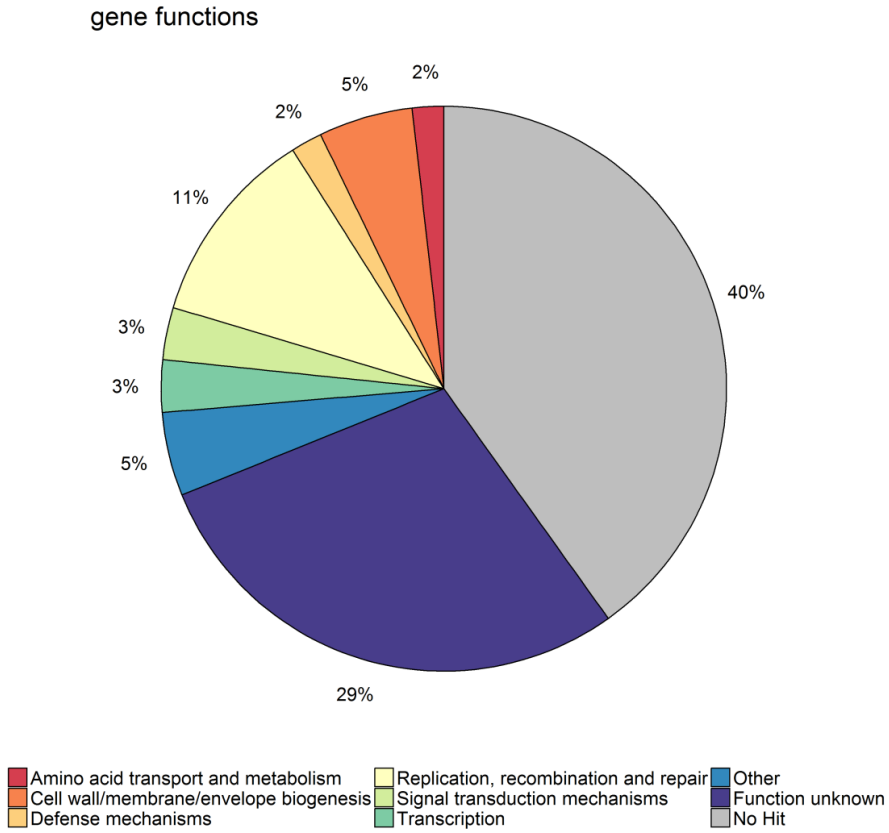


Figure 5. Functional annotation categories of the eggNOGs matching to the set of significant genes across all nine phages.

To estimate whether this observed distributions of functional categories in the 167 significant gene families is different from what could be expected by chance, we estimated cumulative density functions for each eggNOG category from 10.000 random subsamples of 167 gene families drawn from the total set of 6419. From those, we calculated the likelihoods of observing each category by chance, and next evaluated if the probability of a given functional category estimated from the 167 significant gene families is enriched or depleted compared to these random likelihood values.

With a threshold of p=0.05, we find that categories 'No hit' and 'Replication, recombination and repair' are enriched, while 'Post-translational modification, protein turnover, and chaperones' and 'Inorganic ion transport and metabolism' are depleted, see Supplementary Table S3. Further, it is conceivable that many gene families influencing the susceptibility are themselves phage-associated, as is evidenced in the functional annotation terms found for them. As phage genomes typically suffer from poor annotation [15], it is not surprising to find a high percentage of gene families without hits to the database and with hits to the 'unknown function'.

2.6 Overlap of significant gene family sets

We further analyzed the overlap between the significant gene family sets found for each phage model. Figure 6 shows a histogram of the number of phage models where a given gene family was identified significant. It clearly presents that very few significant gene families are shared by many phage models and only one is shared by all nine. The majority of significant gene families have been observed in interaction with only one or two different phages. This in turn means that each of the phages we tested has a distinct and specific interaction with our bacterial strain set, since different genes in the bacterial host dictate whether infection will be successful.

Further, the significant gene families of the three cocktails are not a linear combination of the sets identified for their component phages though there is a sizeable overlap (data not shown).

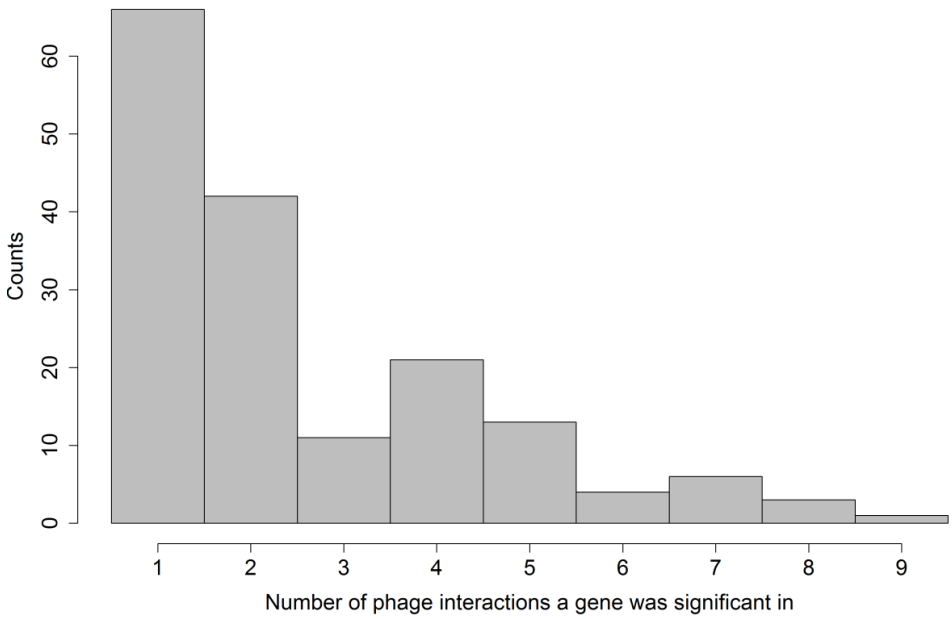


Figure 6. Histogram depicting the number of phage models where a given gene family was identified significant.

There were four gene families found significant in at least eight phage models. They are listed in Table 4 along with their direction of influence and the annotation and category of their matching eggNOG, if any. Out of the four, three increase resistance to phage while one was ambiguous in its direction of influence. Two gene families had no hit in the eggNOG database and one was categorized as being of 'unknown function'. We were therefore unable to deduce a possible function for them though they appear to be of great importance for phage susceptibility. One, cluster 3112, appears to be involved in regulation of transcription and signal transduction which may play a role in host takeover. There were no direct indications for how exactly those gene families effect their influence biologically but it is evident from the models that they do.

Table 4. Predicted functions of the gene families found significant in interaction with eight or more phages.

Gene family ID	Times observed	Increases	eggnog annotation	eggnog category
cluster_1791	9	Resistance	-	No Hit
cluster_389	8	Resistance	-	Function unknown
cluster_3112	8	Resistance	Transcriptional regulator	Transcription
cluster_3992	8	Ambiguous*	-	No Hit

*This gene family always confers phage resistance except in one interaction in which it confers susceptibility.

3. Discussion

In this study we sought to model the host-genetic determinants of MRSA phage susceptibility with a two-step logistic regression model fitted via ridge regression. We succeeded in building models of acceptable performance for nine of the 12 tested phage preparations with AUCs ranging from 0.65 to 0.87. By doing so, we identified 167 host gene families that influence *S. aureus*' interaction with those nine phages.

Our dataset is with 207 observations rather small for this type of analysis, since there are many more covariates, i.e. gene families than observations.

We have addressed this by building a two-step model and including a filtering step based on p-values, thereby greatly reducing the number of covariates going into the analysis. As biological entities are shaped by evolution, the strains share some degree of relatedness, and the testing results are not completely independent observations. We have partitioned the data according to phylogeny in a way that ensures highly similar strains are located to the same partition. Doing that ensures that the observations we are aiming to predict are more independent from the ones we feed into the model during training. The partitioning was maintained at all steps, ensuring that data from highly similar strains was never used to predict the outcome.

Furthermore, there was an uneven partitioning of the data due to a high percentage of strains from two very related sequence types, which may lead to bias. The challenge of uneven partitions was addressed by subsampling the oversized partition 1 so we could obtain a realistic distribution of p-values for the association of all genes to the observed phenotype. Lastly, our set of strains with its composition of clonal complexes is specific to Denmark [16] . It is not necessarily representative of *S. aureus* populations observed in different settings.

It should further be noted that our approach can only identify gene families that are part of the accessory genome, since the first selection step is based on differential abundance of those gene families in susceptible vs resistant strains. Furthermore, this analysis does not consider point mutations as far wild type and mutant version of a gene are more than 90% identical, since we have clustered genes into families with that threshold.

Regarding the electronic gene family annotation, we were able to identify four gene families related to restriction-modification systems, all of which increased the resistance to phage as expected.

Further, six of the significant gene families were related to transcriptional regulation and a multitude of gene families appear to be mobile elements of some kind. Those gene families had varying direction of influence. The findings fit well with the fact that phages try to shut down host transcription during take over, as well as with the interplay of integrated prophages and external phages, which can either complement each other or oppose each other. An integrated prophage may for example protect from further infection via a principle known as superinfection-exclusion [17].

For a large proportion of the significant gene families however, no hit could be found in the eggNOG database and of those that had a hit, the most common category was 'Function unknown'. This may be due to the fact *S. aureus* has a large accessory genome that is made up mostly of different types of mobile genetic elements, among them prophages, highly diverse and not well characterized [8].

We also found that there is only a minor overlap between the sets of significant gene families identified for different phages. This means that each phage had a different and specific interaction with the set of bacterial strains.

Further, we found that more gene families promoted resistance than susceptibility. Among the four gene families that were found significant in interaction with at least eight different phages, three promote resistance and one was ambiguous (see Table 4). This overrepresentation of gene families promoting resistance was expected, since in our set-up resistance to phage can more easily be explained by a gain of function model, meaning the gaining of a defense mechanism of which there are plenty found in nature. We were unfortunately unable to identify the nature of the defense mechanism in most resistance promoting gene families from electronic annotation alone.

Conversely, a gain in susceptibility linked to the presence of a certain gene family is more difficult to explain. The most ready interpretation is that this gene family somehow improves conditions for the phage. The observation can also be explained by integrated prophages which may become activated upon infection or stress caused by the adsorption of an external phage and then lyse their host after completing the lytic cycle. Since the products of the bacterial lysis by the phages were not sequenced, we cannot say whether the external, therapeutic phage or an integrated prophage is the agent of the lysis. Intriguingly, evidence of an interplay between virulence and phage resistance has also been shown. Laanto *et al* report that after co-cultivation with lytic phage, strains of the fish pathogen *Flavobacterium columnare* that have acquired phage-resistance have also lost their virulence compared to phage-sensitive paternal strains [18]. Similar observations have been made for *S. aureus* by Capparelli *et al* [19] , who show that phage-resistance is associated with reduced fitness. Accordingly, genes families found by us to increase phage susceptibility may also be associated with virulence and competitiveness. This is coherent with the origin of our strain set as clinical patient isolates.

One of the current debates in phage therapy focuses on the issue of whether broad spectrum phage cocktails or monovalent phage preparations are preferable [13], [20] . Our approach is a step in the direction of characterizing the interplay between clinical strains of MRSA and single phage

preparations so that a well-targeted phage can be utilized for therapy. We have not observed an advantage of phage cocktails over the monovalent preparations they contain. This may be due to interference between the component phages as has for example been documented by Delbruck [21] and Adams [22].

We have shown that while our methodology does not have predictive power, allow for the association of the observed phenotype with the genetic background, thereby producing interpretable results that can be used for gene function discovery. This type of analysis, which combines phenotypic and whole genome sequencing (WGS) data can be used to identify genetic determinants of observed bacterial phenotypes in other settings as well.

4. Materials and Methods

4.1 Collection of clinical MRSA strains used for susceptibility testing

The collection of 207 MRSA strains tested in this project as well as their whole genome sequences (WGS) were obtained from the Clinical Microbiology Department of Hvidovre Hospital, Denmark. The strains originate from patient samples. They were selected to represent a broad genetic diversity of the more than 5000 WGS MRSA from Hvidovre Hospital.

Although no methicillin-sensitive (MSSA) strains were included in the study, we nonetheless chose MRSA strains of the spa-types that are common in MSSA infections. We included MRSA strains positive for PVL and containing *mecC*. All inclusion criteria are listed in a Supplementary file and the properties of selected isolates can be found in the Supplementary Table S1.

4.2 Collection of phages used for susceptibility testing

A total of 12 therapeutic staphylococcal phage preparations were used for susceptibility testing. They contain phages which are part of the proprietary collection of therapeutic phages used by the phage therapy unit of the Hirszfeld Institute of Immunology and Experimental Therapy of the Polish Academy of Science in Wroclaw (HI) [23]. Nine of the preparations are monovalent phage lysates (containing 1N/80, 676/F, 676/T, 676/Z, A3/R, A5/L, A5/80, P4/6409, or phi200/6409 phage). Crude phage lysates were prepared according to the modified method of Ślopek *et al*. [9] [citation]. Six of those phages (1N/80, 676/Z, A3/R, A5/80, P4/6409, and phi200/6409) were sequenced and confirmed to be obligatory lytic and belonging to a *Twortlikevirus* genus of a *Spounavirinae* subfamily of *Myoviruses* [24]. They were provided in routine test dilution (RTD) which is the highest dilution that still gives confluent lysis on the designated propagating strain of *S. aureus* [25]. Three others were equal mixtures of A5/80, P4/6409, and 676/Z phages prepared at the Institute of Biotechnology, Sera and Vaccines BIOMED S.A. in Cracow, Poland: MS-1 phage cocktail lysate containing each phage in a titer no less than 5×10^5 pfu/ml, OP_MS-1_TOP cocktail of purified phages suspended in phosphate buffered saline containing each phage at no less than 10^9 pfu/ml [26] [citation], and OP_MS-1 phage cocktail of the similar characteristics as OP_MS-1_TOP but containing up to 10% of saccharose as a phage stabilizer.

4.3 Susceptibility testing procedure

Testing for phage susceptibility was performed as described by Ślopek et al [27]. In short, 50 µl of phage preparation was applied onto a fresh bacterial lawn from day culture and the results were assessed the next day following 6 hours incubation at 37°C.

Results were assessed according to a 7-point scale as described by Ślopek et al [27] and then further discretized into two levels: 'susceptible' and 'resistant'. The 'susceptible' label was applied to the two strongest reactions, resulting in confluent or semi confluent lysis. According to standards

applied at the Bacteriophage Laboratory of the HI, those two levels enable the phage procurement for therapeutic phage preparation. All other weak reactions as well as a negative reaction and opaque lysis were regarded as 'resistant'. The full set of 207 strains was challenged with each of the 12 phage preparations. We call the result of susceptibility testing to a preparation the 'interaction' of our strain set with said phage.

We also build models using a modified division of the phage reaction including weakly susceptible levels (>20 independent plaques) in the definition of 'susceptible' and only including strongly resistant results (resulting in the negative reaction, opaque lysis or < 20 plaques) in the 'resistant' label. Thereby, we investigated whether the split we imposed on the 7-scale phage typing results influenced our modeling results.

4.4 Data Partitioning

For the purpose of modelling the phage response from the genomic composition of the bacterial strains, the 207 MRSA strains were divided into five partitions. This division was based on the orthogonal average nucleotide identity (orthoANI) as described by Lee et al [28]. OrthoANI is suitable for creating a distance matrix, because it is a symmetric measure of distance, unlike the traditional ANI. Calculations were performed on all pairs of strains with the standalone tool OAT by Lee et al. Distances were subsequently calculated as 1 – orthoANI and a heat map was generated which can be found in Figure 1.

The resulting heat map showed very clear clusters of closely-related sequences. Partitioning was therefore done by visual inspection.

The partitions thus obtained were then used in a five-fold cross validation framework, i.e. four of them were combined into the training set and one was left out for testing. This process is repeated five times so that each partition is in turn the testing set.

4.5 Identification of gene families

The genetic makeup of the MRSA strains was analyzed by first predicting genes and performing functional annotation through the RAST service [29]. The predicted genes were then clustered with cd-hit [30] using a cutoff of 90% on global sequence identity, word size 5 and the -g 1 option to cluster with the best match instead of the first match. This resulted in a total of 6.419 gene families in the 207 MRSA strains.

Next, the feature space, i.e. the number of gene families included, was reduced by removing gene families with limited power for distinguishing susceptible from non-susceptible bacterial strains. This was done by constructing 2x2 contingency tables as shown in Table 3, and from these tables calculating a p-value to each gene family in each phage interaction using Fischer-Boschloo's exact unconditional test. In contrast to the often used Fischer's, exact conditional test, Fischer-Boschloo's is an exact unconditional test. In total sum fixed designs, unconditional test are always preferable to conditional tests for reasons detailed by Lydersen et al [31]. We then imposed a threshold of 0.01 on the p-value for the gene family to be admitted to the second step of modelling.

Table 3. Layout of the contingency tables used for analysis. The asterisk denotes the total sum fixed by design.

	Susceptibility		
Presence of gene family	susceptible	resistant	Sum
present	a[1]	b[2]	a+b
absent	c[3]	d[4]	c+d
Sum	a+c	b+d	n*

[1] Number of isolates that are susceptible to the phage currently looked at and in which the current cluster is present.
[2] Number of isolates that are resistant to the phage currently looked at and in which the current cluster is present.
[3] Number of isolates that are susceptible to the phage currently looked at and in which the current cluster is absent.
[4] Number of isolates that are resistant to the phage currently looked at and in which the current cluster is absent.
Both the row and column margins sum to n.

This first filtering step was performed inside the cross validation framework, so that the partition being tested was not included in this initial p-value based feature reduction. Due to the fact that the 2x2 tables were constructed from only the training set, some gene families in a given test set do not have a p-value associated. This specific situation arises when gene families are only present in one partition and that partition is left out of the training set. In these cases the gene family was assigned a p-value of NA (not applicable).

4.6 Bootstrapping

As can be seen in Figure 1, partition 1 is substantially larger than the other four partitions in the benchmark data, see 3.2 for further details. This potentially imposes a bias when calculating the association p-values, since these often will be driven solely by the data in partition 1. To amend that, a bootstrapping resampling procedure was applied to partition 1: When picking gene families based on a combination of partitions that includes partition 1, instead of including the full partition, a subsample of 25 strains used and was added to the other three partitions. From that data, we then created contingency tables and calculated p-values as described above. This procedure was repeated a 100 times, resulting in 100 p-values per gene family per phage interaction. We then imposed the condition that a gene family had to pass the p-value threshold of 0.01 in at least 90 of those to be selected.

4.7 Model construction and feature selection

While a strong p-value obtained, for instance, from a contingency table as described above is an indication, it is often not a conclusive proof of an actual association existing between the gene and the observed phenotype. For that, the gene needs to have predictive power towards the phenotype it is thought to be influencing. Therefore, we chose to model the phage response with a logistic regression model that was fitted using a Ridge regression.

For each phage interaction, a logistic Ridge regression model was trained on four of the partitions and tested on the one left out partition using the gene families that has passed the association-based p-values criteria described above as input, and the binary susceptible/resistant annotations as target values. This was done five times for the five possible combinations of partitions. This five-fold cross validation framework allowed us to evaluate the model's predictive potential and assess their robustness.

In this way, five models were constructed each with regression weights associated to each of the gene families. If a gene family had not been picked for that particular partition, it was assigned a weight of NA. We hypothesised that gene families with a high weight across many partitions drive the response to this particular phage. In order to verify this, we next trained and tested a second five-fold cross validated regression model with only the genes that 1) were significant according to the Fischer-Boschloo's test ($P \leq 0.01$) and 2) had weights above 0.01 in at least three partitions in the first regression model.

In order to verify that the set of genes we identified were indeed descriptive of the phage susceptibility and not an artifact of over-fitting, we repeated the model construction and feature selection with shuffled target values. That is, we randomly associated susceptibility outcomes and bacterial genomes, while keeping the ratio between susceptible and resistant as in the original data. We then re-ran the modelling, and evaluated the predictive performance and the number of predictive gene-families identified.

4.8 Assignment of EggNOGs

We further compared each gene family to the EggNOG database [32] by using the eggNog-mapper available on their webpage. EggNOG is a database of non-supervised orthologous groups (NOG) of proteins, in which each group has only one annotation term compiled from the integrated and summarized functional annotation of its group members. Each NOG is also part of a broader functional category. This allows for the quick and efficient assignment of functions for predicted genes by finding their matching NOG.

After identifying a set of significant gene families (see 2.7), the prevalence of each functional category in that set was calculated. We also extracted 10.000 random subsamples of the same size from the full set of genes and used these data to establish an estimated cumulative density function (eCDF) for the prevalence of each category.

Supplementary Materials: The following are available online at www.mdpi.com/link: Details of inclusion criteria for MRSA strains. Figure S1: Plot of the cumulative mean square error of the inner cross validation vs strength of the ridge penalty. Figure S2: P-value distributions of gene enrichment analysis on phage preparations 1N_80, A3_R and cocktail MS-1. Table S1: List of MRSA strains included in the test set. Table S2: List of all significant gene families along with their functional annotation terms. Table S3: Probabilities of observing a given prevalence per functional category based on the cumulative density function. Table S4: Detailed phage typing results.

Acknowledgments:

This work was supported financially by a full PhD scholarship granted by the Technical University of Denmark (DTU).

Author Contributions: Mette V. Larsen and Ryszard Międzybrodzki conceived and designed the overall project idea. Morten Nielsen coordinated the modeling part. Mette V. Larsen and Morten Nielsen coordinated the gene functional analysis. Ryszard Międzybrodzki and Ewa Jończyk-Matysiak coordinated the experimental part. Ewa Jończyk-Matysiak and Henrike Zschach conducted the laboratory work. Beata Weber-Dąbrowska supplied the phage preparations. Henrik Westh supplied the bacterial strains and advised on the strain selection criteria. Henrik Hasman, Henrik Westh and Andrzej Górski provided feedback on the biological relevance of the findings. Henrike Zschach and Ryszard Międzybrodzki wrote the paper. Mette V. Larsen, Morten Nielsen and Andrzej Górski advised the paper writing and performed edits. All authors contributed to the final proof read.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results

References

[1] WHO, "Antimicrobial resistance fact sheet," 2016. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs194/en/. [Accessed: 05-Sep-2017].

[2] WHO, "WHO (2017) Global priority list of antibiotic-resistant bacteria," *27.02.2017*, 2017. [Online]. Available: http://www.who.int/medicines/publications/WHO-PPL-Short_Summary_25Feb-ET_NM_WHO.pdf.

[3] S. Chhibber, T. Kaur, S. Sandeep Kaur, B. Wilson, and A. Cheung, "Co-Therapy Using Lytic Bacteriophage and Linezolid: Effective Treatment in Eliminating Methicillin Resistant Staphylococcus aureus (MRSA) from Diabetic Foot Infections," *PLoS One*, vol. 8, no. 2, p. e56022, Feb. 2013.

[4] S. T. Abedon, S. J. Kuhl, B. G. Blasdel, and E. M. Kutter, "Phage treatment of human infections.," *Bacteriophage*, vol. 1, no. 2, pp. 66–85, Jan. 2011.

[5] N. B. Pincus, J. D. Reckhow, D. Saleem, M. L. Jammeh, S. K. Datta, and I. A. Myles, "Strain specific phage treatment for Staphylococcus aureus infection is influenced by host immunity and site of infection," *PLoS One*, vol. 10, no. 4, p. e0124280, Apr. 2015.

[6] R. Międzybrodzki *et al.*, "Clinical aspects of phage therapy," *Adv Virus Res*, vol. 83, pp. 73–121, 2012.

[7] J. Borysowski, M. Łobocka, R. Międzybrodzki, B. Weber-Dabrowska, and A. Górski, "Potential of Bacteriophages and Their Lysins in the Treatment of MRSA," *BioDrugs*, vol. 25, no. 6, pp. 347–355, Dec. 2011.

[8] M. Deghorain and L. Van Melderen, "The Staphylococci Phages Family: An Overview," *Viruses*, vol. 4, no. 12, pp. 3316–3335, Nov. 2012.

[9] S. Ślopek, I. Durlakowa, B. Weber-Dąbrowska, A. Kucharewicz-Krukowska, M. Dąbrowski, and R. Bisikiewicz, "Results of bacteriophage treatment of suppurative bacterial infections. I. General evaluation of the results.," *Arch. Immunol. Ther. Exp*, vol. 31, pp. 267–291, 1983.

[10] B. Weber-Dąbrowska, E. Jończyk-Matysiak, M. Żaczek, M. Łobocka, M. Łusiak-Szelachowska, and A. Górski, "Bacteriophage Procurement for Therapeutic Purposes," *Front. Microbiol.*, vol. 7, p. 1177, Aug. 2016.

[11] J. E. Samson, A. H. Magadán, M. Sabri, and S. Moineau, "Revenge of the phages: defeating bacterial defences.," *Nat. Rev. Microbiol.*, vol. 11, no. 10, pp. 675–87, Oct. 2013.

[12] G. Xia and C. Wolz, "Phages of Staphylococcus aureus and their impact on host evolution.," *Infect. Genet. Evol.*, vol. 21, pp. 593–601, Jan. 2014.

[13] J. P. Pirnay *et al.*, "The phage therapy paradigm: Prêt-à-porter or sur-mesure?," *Pharm. Res.*, vol. 28, no. 4, pp. 934–937, 2011.

[14] S. Monecke *et al.*, "A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant Staphylococcus aureus.," *PLoS One*, vol. 6, no. 4, p. e17936, Jan. 2011.

[15] B. L. Hurwitz, J. M. U'Ren, and K. Youens-Clark, "Computational prospecting the great viral unknown," *FEMS Microbiol. Lett.*, vol. 363, no. 10, May 2016.

[16] M. Bartels *et al.*, "Monitoring meticillin resistant Staphylococcus aureus and its spread in Copenhagen, Denmark, 2013, through routine whole genome sequencing," *Eurosurveillance*, vol. 20, no. 17, p. 21112, Apr. 2015.

[17] B. Hofer, M. Ruge, and B. Dreiseikelmann, "The superinfection exclusion gene (sieA) of bacteriophage P22: Identification and overexpression of the gene and localization of the gene product," *J. Bacteriol.*, vol. 177, no. 11, pp. 3080–3086, 1995.

[18] E. Laanto, J. K. H. Bamford, J. Laakso, and L. R. Sundberg, "Phage-Driven Loss of Virulence in a Fish Pathogenic Bacterium," *PLoS One*, vol. 7, no. 12, 2012.

[19] R. Capparelli *et al.*, "Bacteriophage-resistant Staphylococcus aureus mutant confers broad immunity against staphylococcal infection in mice.," *PLoS One*, vol. 5, no. 7, p. e11720, Jan. 2010.

[20] A. Górski *et al.*, "Phage therapy: Combating infections with potential for evolving from merely a treatment for complications to targeting diseases," *Front. Microbiol.*, vol. 7, no. SEP, pp. 1–9, 2016.

[21] D. M, "Interference Between Bacterial Viruses: III. The Mutual Exclusion Effect and the Depressor Effect.," *J. Bacteriol.*, vol. 50, pp. 151–170, 1945.

[22] M. H. Adams, *Bacteriophages*. New York: Interscience Publishers, 1959.

[23] B. Weber-Dąbrowska, M. Mulczyk, A. Górski, J. Boratyński, M. Łusiak-Szelachowska, and D. Syper, "Methods of polyvalent bacteriophage preparation for the treatment of bacterial infections.," US Patent No. US 7232564 B2, 2002.

[24] M. Łobocka *et al.*, "Genomics of Staphylococcal Twort-like Phages - Potential Therapeutics of the Post-Antibiotic Era," *Adv. Virus. Res.*, vol. 83, pp. 143–216, 2012.

[25] J. E. Blair and R. E. Wiliams, "Phage typing of staphylococci," *Bull World Heal. Organ.*, vol. 24(6), pp. 771–84, 1961.

[26] A. Górski, B. Weber-Dąbrowska, R. Miedzybrodzki, G. Stefański, K. Dechnik, and E. Olchawa, "A method for obtaining bacteriophage purified preparations.," Polish patent No.PL 212811 B1, 2012.

[27] S. Slopek, I. Durlakowa, A. Kucharewicz-Krukowska, T. Krzywy, A. Slopek, and B. Weber, "Phage typing of Shigella flexneri," *Arch. Immunol. Ther. Exp.*, vol. 20, no. 1, 1972.

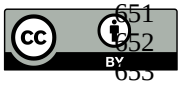
[28] I. Lee, Y. O. Kim, S. C. Park, and J. Chun, "OrthoANI: An improved algorithm and software for calculating average nucleotide identity," *Int. J. Syst. Evol. Microbiol.*, vol. 66, no. 2, pp. 1100–1103, 2016.

[29] R. Overbeek *et al.*, "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D206-14, Jan. 2014.

[30] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.

[31] S. Lydersen, M. W. Fagerland, and P. Laake, "Recommended tests for association in 2 x 2 tables.," *Stat. Med.*, vol. 28, no. 7, pp. 1159–75, Mar. 2009.

[32] J. Huerta-Cepas *et al.*, "eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D286–D293, Jan. 2016.



Part III

Conclusion

10 Conclusion and outlook

The vast increase of antimicrobial resistance seen in recent years poses a serious threat to public health and - if unresolved - may lead to a future where common bacterial infections will once again be deadly. Phage therapy is one of the most promising alternatives to antibiotics and accordingly considerable time and effort have been invested into the field.

The focus of my PhD has been to investigate how genomics and machine learning techniques can be used to further the understanding of therapeutic phages and the phage-host interaction. To do so, different aspects of phage therapy research were explored. It started with the characterization of an already existing phage cocktail, then moved on to investigate determinants of phage susceptibility in the host-genome. Lastly, the diversity of phages present in sewage, the major isolation source of therapeutic phages, was explored. To maintain the flow of thought, the chronologically last project (concerning phages in sewage) was presented in this thesis as the second project.

In the first project of this PhD, the long-used and highly clinically relevant INTESTI phage cocktail has been sequenced and analyzed. We found that there are at least 23 different phage types in the cocktail, 20 of which showed considerable similarity to known phages while 3 were largely novel.

One of the main conclusions of this paper was, that the different phage types were present in vastly different abundances in the cocktail. This could be a consequence of the way the cocktail is produced. However, since the INTESTI cocktail has been in use successfully for many years, the uneven composition may also be a feature. Different phage types exhibit different levels of stability, efficiency of adsorption and burst size. Some phages may therefore require a higher multiplicity of infection that others to be effective. This has implications for companies and research laboratories seeking to produce their own cocktail. It should be verified whether even or uneven ratios of the component phages produce the best results.

The study furthermore included an amplification experiment in which in-house bacterial strains that had proven susceptible to the cocktail were infected and subsequently their lysates sequenced to discover which phages in the cocktail had been amplified. It was found that a phage contig which was barely present in the sequencing data of the full cocktail corresponded to a

group of contigs with high coverage in the lysate of *Pseudomonas Aeruginosa* PAO1. This result illustrates the usefulness of a highly diverse cocktail since even component phages in low abundance can unfold their potential upon meeting their host. However, in current phage therapy efforts in the Western world, phage cocktails are typically of low complexity because drug regulatory authorities require approval for every component phage. To resolve this question, there is a need for more studies investigating whether or not high-complexity phage cocktails are preferable from a clinical point of view.

In the second project, phage communities present in sewage samples were compared to known phages in databases as well as to each other. It was found that typically more than 50% of phage contigs in a sample had no close known relative. This underlines both a great need for more environmental phage studies as well as the enormous genetic potential still hidden in even mundane environments like city sewage.

The study also showed that the phage communities of different sewage samples were quite distinct from each other. This pairwise genomic distance, based on shared k-mers, was astonishingly constant and did not appear to correlate with whether or not the samples were from a similar geographic location. At the same time, the majority of samples contained crAssphage, a highly abundant phage in human fecal metagenomes [58]. Both of these observations suggest that there may be principles underlying the phage community in sewage that are invariant to geographic location.

When looking into how well known phages are represented in the sewage samples, we furthermore observed intriguing patters in the ANI distributions. Two of those seemed to correlate with the *Siphoviridae* and *Myoviridae* families. Though we do not currently understand the meaning of these patterns, it would be of interest to investigate whether they could be related to some property of the phage family, such as a preferential mode of mutation, or to see if they also hold for phage samples from other environments.

Finally, the third project was centered on identifying gene families in the accessory genome of the pathogen *S. aureus* that influence its susceptibility to phages. For this, 207 strains of MRSA were tested for susceptibility to 12 different phage preparations. As a result, 167 such gene families were found by building nine successful regression models. Among those were genes related to prophages and mobile elements, restriction-modification systems and transcriptional regulators. However, most of the identified gene families were of unknown function. This illustrates another aspect of the phage susceptibility problem: Though *S. aureus* is an important pathogen, large parts of its accessory genome remain poorly characterized. To better understand the phage-host interaction it will be vital to assign functions to a larger part of

the host genome.

This project further showed that most gene families were only found to influence susceptibility to some of the nine phage preparations. In other words, each of the nine phage preparations had a distinct and specific interaction with the strain set. This result reinforces the notion that there is an array of different phage defense mechanisms, at least in *S. aureus*, and not one way to gain resistance to the majority of phages. A next step in this line of questioning could be to experimentally verify whether the identified gene families are actually causal of phage resistance/susceptibility, for example via knock-out experiments. There may however be significant challenges to that because it might actually be a combination of gene families that is causing the phenotype. Therefore, a large number of experiments may be necessary.

Phage susceptibility is also not purely genetically determined. Environmental factors can have an influence, for example via an up- or down regulation of receptor expression. Høyland-Kroghsbo *et al* describe such a phage defense mechanism based on quorum sensing in *E. coli*, where a phage receptor is down-regulated based on population density [16]. Those effects are not currently captured in our model, but could be included in future studies.

Understanding the genetic determinants of susceptibility is an important step forward towards evidence-based selection of the appropriate therapeutic phage preparation. This ties in with a general movement towards personalization in medicine. It can furthermore aid in the rational design of phage cocktails by combining phages for which different sets of resistance-promoting gene families have been identified. This would indicate that those phages have different modes of action or at least cannot be evaded with the same strategy. Using such complementary phages could then delay the development of phage resistance in the bacterial population.

Looking at the broader picture, the first and third project of this PhD have dealt with the two principle approaches to phage therapy: Using either standardized, off-the-shelf cocktails or personalized phage preparations tailored to the infecting strain. Each of these has merits as well as drawbacks. First-off, a customized phage preparation is more sure to eliminate specifically the bacterial strain causing the infection and the effect on the commensal microbiome is minimized. However, this advantage comes with the drawback of needing to identify the infecting strain prior to treatment. Further, as pointed out by Pirnay *et al* in their 2011 commentary, custom-made phage preparations are not compatible with the current regulatory guidelines as there is not the time or funding to gain approval for their use through the usual channels, i.e. clinical trials [9].

Ready-made cocktails on other hand have been shown by for example the Eliava Institute to lose efficacy over time and need to be updated by either adapting the component phages to the current bacterial strains or isolating new phages [9]. This again creates problems with the current legislation. In time, guidelines for updating existing cocktails may be set-up, possibly from a library of approved phages. For now, it is not clear which criteria would need to be met and which sort of characterization to be provided for a new phage to be added into an approved product. For more information concerning this subject see [59] for the publicly available transcripts of the FDA workshop on 'Bacteriophage Therapy: Scientific and Regulatory Issues' in July 2017.

There is no final verdict on this question. However, those two approaches need not be exclusive of each other. They could also be used in tandem as each is suited for different purposes. Ready to use cocktails could be applied as a first line drug and for prophylactic purposes in wound care, as is done in the Republic of Georgia, where phage preparations are part of the standard medical care [60]. Custom-made phage preparations could be used for complicated infections and cases where the standard cocktails have proven ineffective.

In conclusion, after waiting in the wings for many years the time is now right for phages to takes center stage once more and become an integral part in combating bacterial infection in Western Medicine. The work presented in this thesis is a step in the direction of bringing the field of phage research further towards a future of phage therapy in humans.

Part IV

Appendix

A Supplementary Material for Paper I

Table A1. Overview of the trimming parameters and assembler that gave the best result for each phage DNA sample. Trimming was based on the output of FASTQC

Sample	Trimming	Assembler
E. coli	removed nucleotides from the right (5' end) according to quality score (min 20) removed reads according to the mean quality (min 20) removed reads shorter than 50 bp remove reads with streaks of N longer than 10 removed 20 nucleotides on the left (3' end) removed 10 nucleotides on the right (5' end) removed duplicate reads	Genovo
Enterococcus	removed nucleotides from the right (5' end) according to quality score (min 20) removed reads according to the mean quality (min 20) removed reads shorter than 50 bp remove reads with streaks of N longer than 10 removed 30 nucleotides on the left (3' end) removed 10 nucleotides on the right (5' end) removed duplicate reads	Genovo
P. aeruginosa PAO1	removed nucleotides from the right (5' end) according to quality score (min 20) removed reads according to the mean quality (min 20) removed reads shorter than 50 bp remove reads with streaks of N longer than 10 removed 20 nucleotides on the left (3' end) removed 50 nucleotides on the right (5' end) removed duplicate reads	Genovo
P. aeruginosa 0407431-2	Untrimmed	Velvet
Proteus	Untrimmed	Velvet
Salmonella	removed nucleotides from the right (5' end) according to quality score (min 20) removed reads according to the mean quality (min 20) removed reads shorter than 36 bp remove reads with streaks of N longer than 10	Genovo
Shigella flexneri	removed nucleotides from the right (5' end) according to quality score (min 20) removed reads according to the mean quality (min 20) removed reads shorter than 50 bp remove reads with streaks of N longer than 10 removed 20 nucleotides on the left (3' end) removed 10 nucleotides on the right (5' end) removed duplicate reads	Genovo
Shigella sonnei	removed nucleotides from the right (5' end) according to quality score (min 20) removed reads according to the mean quality (min 20) removed reads shorter than 50 bp remove reads with streaks of N longer than 10 removed 20 nucleotides on the left (3' end) removed 10 nucleotides on the right (5' end) removed duplicate reads	Genovo

Table A2. Overview of the bacterial strains used for small scale susceptibility testing. Observe that all *Salmonella* are of the species *Salmonella enterica* subsp. *enterica* but they are identified as different serovars. Reference strains are marked with an asterisk. Pathogenic strains are marked with a plus, opportunistic pathogens with a tilde. All strains are part of an in-house collection.

Genus	Species/ Serovar	Strain	Susceptibility
Salmonella	serovar Enteritidis	ATCC 13076 *+	Yes
	serovar Typhimurium	ATCC 14028 *+	Yes
	serovar Saint Paul	DVL31 +	Yes
	serovar Newport	EQAS1 98-24475-1+	Yes
	serovar Infantis	EQAS1 98-74091-5+	Yes
	serovar Derby	EQAS2 99-65209-5+	Yes
	serovar Typhimurium	DT36 +	Yes
	serovar Enteritidis	PT1 +	Yes
	serovar Heidelberg	75-12893-1+	Yes
	serovar Dublin	1111H11036 +	Yes
Staphylococcus	*aureus*	ATCC 29213 *+	No
	aureus	ATCC 25923 *+	Yes
	epidermidis	CCM2354	No
	pseudointermedius	Bjorn 55-4	No
	hyicus	NCTC 10350	No
	felis	Sneleopard	Yes
	lugdunensis	E2-1928945	No
	aureus	76670 CC8 related +	Yes
	aureus	Not given+	Yes
	aureus	MSSA A7+	Yes
Shigella	*flexneri*	1s +	Yes
	sonnei	2s +	Yes
	boydii	Not given+	Yes
	flexneri	Not given+	Yes
	Not given	HN-Sh, 2006-001, 2007-5-3 +	Yes
Pseudomonas	*aeruginosa*	DMS 1128 / ATCC9027 *~	No
	aeruginosa	Skejby_2~	No
	aeruginosa	07 52277-1~	Yes
	aeruginosa	PAOI seq ~	Yes
	aeruginosa	0173267-5~	Yes
	aeruginosa	0407431-2~	Yes
	aeruginosa	0107338-1~	Yes
Escherichia	*coli*	ATCC 25922 *	Yes
	coli	C 64-12 +	No
	coli	C 60-12 +	No
	coli	C 23-12 +	No
	coli	oedemsyge-45	No

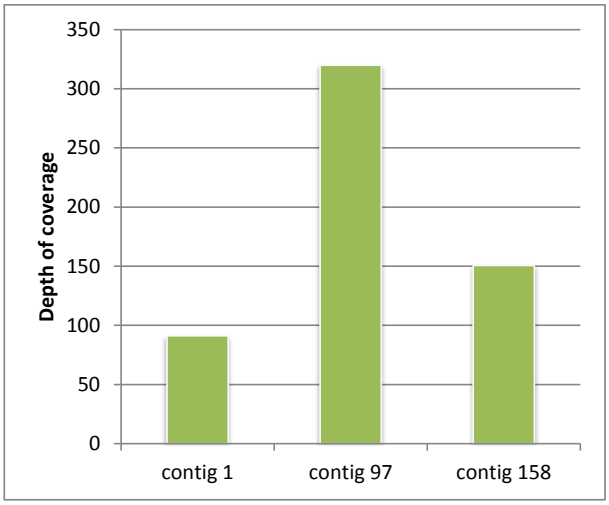
	coli	BW25II3	Yes
Proteus	*hauseri*	DSM 30118/ ATCC 13315 *~	Yes
	vulgaris	DMS 2140/ ATCC 8427 *~	No
	vulgaris	CCUG 36761, ATCC 13315 *~	Yes
	mirabilis	76499961~	Yes
	mirabilis	E2 1928244~	No
Enterococcus	*faecalis*	2011-70-7-6 to 2011-70-250-4 ~	No
	faecium	2011-70-7-8 to 2011-70-252-10~	Yes
	faecalis	2008-37857~	No
	faecalis	12 E ~	No
	faecalis	ATCC 29212 *~	Yes

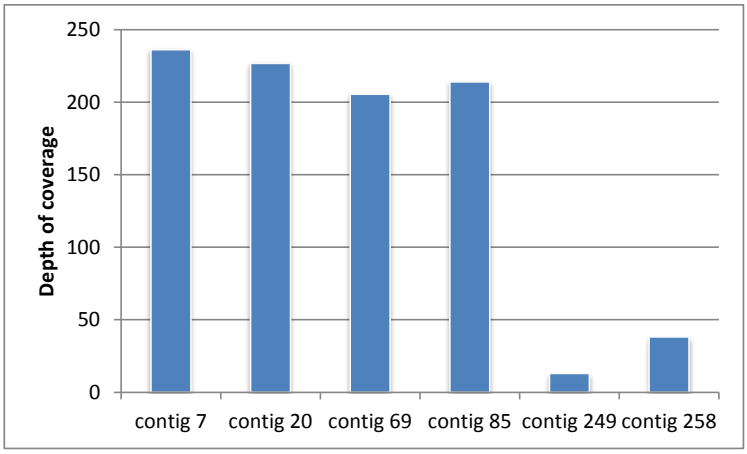
Table A3. Overview of phage clusters identified in the sequencing data of the host-amplified samples. Note that many clusters are much smaller in size compared to the corresponding clusters in the full cocktail. Those clusters have likely not been amplified by that particular host. Some clusters however, e.g. EntF2 and Pao1_new show a great increase in size. This can be explained by the fact that those are infecting clusters (compare Table 4 in the text) which are in higher abundance in the host-amplified samples compared to their original numbers in the cocktail. Therefore, greater parts of those clusters could be recovered from the amplified samples.

Phage Cluster in sample	Cluster size in bp	Corresponding cluster in INTESTI	Size ratio to corresponding cluster
Amplified on *Escherichia coli*			
Eco1	9,737	D1	0.07
Eco2	3,163	D2	0.04
Eco3	19,979	D3	0.23
Eco4	1,043	D4	0.02
Eco5	7,023	D5	0.05
Eco6	133,873	D6	1.64
Eco7	17,744	D7	0.30
Eco9	5,487	D9	0.14
Eco10	39,747	D10	0.27
Eco11	4,131	D11	0.07
Eco12	12,195	D12	0.20
Eco13	9,105	D13	0.05
Eco14	185,358	D14	1.39
Eco15	7,278	D15	0.17
Eco16	18,144	D16	0.39
Eco17	78,630	D17	1.91
Eco18	8,603	D18	0.21
EcoP	41,317	Proteus phage	0.40
Amplified on *Enterococcus faecalis*			
Ent7	58,552	D7	1.01
Ent11	6,268	D11	0.10
Ent13	5,282	D13	0.03
Ent18	41,874	D18	1.02

EntF2	88,702	F2	7.73
Amplified on *Pseudomonas aeruginosa* PAO1_seq			
Pao1_6	9,257	D6	0.11
Pao1_10	1,477	D10	0.01
Pao1_12	538	D12	0.01
Pao1_F1	22,920	F1	1.65
Pao1_P	3,075	Proteus phage	0.03
Pao1_new	45,478	-	19.01
Amplified on *Pseudomonas aeruginosa* 0407431-2			
PA0407_3	87,742	D3	1.00
Amplified on *Salmonella typhimurium*			
Sal3	515	D3	0.01
Sal6	19,359	D6	0.24
Sal7	574	D7	0.01
Sal13	1,047	D13	0.01
Sal14	717	D14	0.01
Sal18	46,366	D18	1.13
SalF2	94,543	F2	8.24
SalP	670	Proteus phage	0.01
Amplified on *Shigella flexneri*			
ShiFl1	2,402	D1	0.02
ShiFl2	4,799	D2	0.06
ShiFl3	1,357	D3	0.02
ShiFl6	21,797	D6	0.27
ShiFl7	3,946	D7	0.07
ShiFl9	3,362	D9	0.08
ShiFl10	1,102	D10	0.01
ShiFl12	7,707	D12	0.13
ShiFl13	1,784	D13	0.01
ShiFl14	177,744	D14	1.34
ShiFl15	48,286	D15	1.10
ShiFl16	4,765	D16	0.10
Amplified on *Shigella sonnei*			
ShiS2	6,868	D2	0.09
ShiS6	11,588	D6	0.14
ShiS14	173,647	D14	1.31
ShiS15	49,031	D15	1.12
ShiS16	4,075	D16	0.09
ShiSP	5715	Proteus phage	0.05
Amplified on *Proteus vulgaris*			
Prot17	59,325	D17	1.44
ProtP	102,963	Proteus phage	0.99

Figure A1. Examples of two clusters who's depth of coverage had a large standard deviation. The lower the contig ID the longer the contig.
Top: Depth of coverage of cluster D1. Contig 1 which is the longest, has a much lower depth of coverage than the short contigs 97 and contig 158. Annotation results showed that many of the genes in contigs 97 and 158 show homology to genes annotated as 'terminal repeat-encoded protein (Tre)'. Bottom: Depth of coverage of cluster D6. The two short contigs 249 and 258 have much lower depth than the other contigs in that group. We theorize that they could represent divergent regions only present in a few of the phages in that cluster.





B Supplementary Material for Paper II

The following supplementary files are available:

1) Supplementary Table S1.
Table of samples including metadata and the amount of phage DNA in base pairs and percent of the full assembly.

Available at:
https://docs.google.com/spreadsheets/d/1c6sIiAbWW6UabYiXkH9cR3mv806DctqkKkIXJYklpKA/edit#gid=0

C Supplementary Material for Paper III

The following supplementary files are available:

1) List of inclusion criteria for MRSA strains.

All strains originate from patient samples. Strains were included in the study if they met one or more of the following criteria:

- Having one of the ten most common spa types that occur in Methicillin-sensitive Staphylococcus aureus infections
- Positive for PVL
- Positive for mecC
- Being of a rare clonal complex
- Being of one of the major clonal complexes prevalent in Europe (cc22, cc30, cc45)
- Being of clonal complex 398 which is typically livestock associated

Additionally, strains where the sequencing data was of good quality were preferred over strains with poor quality sequencing data.

2) Supplementary Table S1.
List of MRSA strains included in the test set and their properties.

Available at:
https://docs.google.com/spreadsheets/d/17ciUDM7rJgmCRjMq-V_xZ23wcd2HrblF206WbtzMst0/edit#gid=0

3) Supplementary Table S2. List of all significant gene families along with the functional annotation terms retrieved from comparison to RAST and eggNOG databases.
A dash ('-') in columns 2-4 indicates that there was no hit found and therefore no annotation term or category could be retrieved. Any other entry is the retrieved annotation term, even if it reads 'NA'. In columns 6-19 'NA' means the gene family was not found significant in that phage model.

Available at:
https://docs.google.com/spreadsheets/d/1joM5QoX5FCE3BI5vPiE3ucFwxDcSvRGR8XuJj82Fn6M/edit#gid=0

4) Supplementary Table S3. Probabilities of observing a given prevalence per functional category based on the cumulative density function. In the first column is noted the observed percentage of genes in a given category, as depicted in Figure 5. The second column shows the probability of

observing this percentage or lower given the estimated CDF. Conversely, the third column shows the probability of observing an even higher percentage given the eCDF.
Note that although categories 'Chromatin structure and dynamics' and 'Extracellular structures' appear overrepresented in the significant gene set via the cumulative density function, this is meaningless since both of categories have been observed zero times in the significant gene set. Those two categories are overall extremely rare within our strain set which makes the cumulative density function collapse.

Letter	**category**	**percent observed**	**p(CDF(x))**	**1-p(CDF(x))**	p<0.05	direction
0	No Hit	40.1%	0.99	0.01	yes	enriched
B	Chromatin structure and dynamics	0.0%	0.98	0.02	yes	enriched
C	Energy production and conversion	0.6%	0.13	0.87	no	
D	Cell cycle control, cell division, chromosome partitioning	1.2%	0.94	0.06	no	
E	Amino acid transport and metabolism	1.8%	0.06	0.94	no	
F	Nucleotide transport and metabolism	0.0%	0.08	0.92	no	
G	Carbohydrate transport and metabolism	1.2%	0.12	0.88	no	
H	Coenzyme transport and metabolism	0.0%	0.06	0.94	no	
I	Lipid transport and metabolism	0.0%	0.12	0.88	no	
J	Translation, ribosomal structure and biogenesis	1.2%	0.17	0.83	no	
K	Transcription	3.0%	0.24	0.76	no	

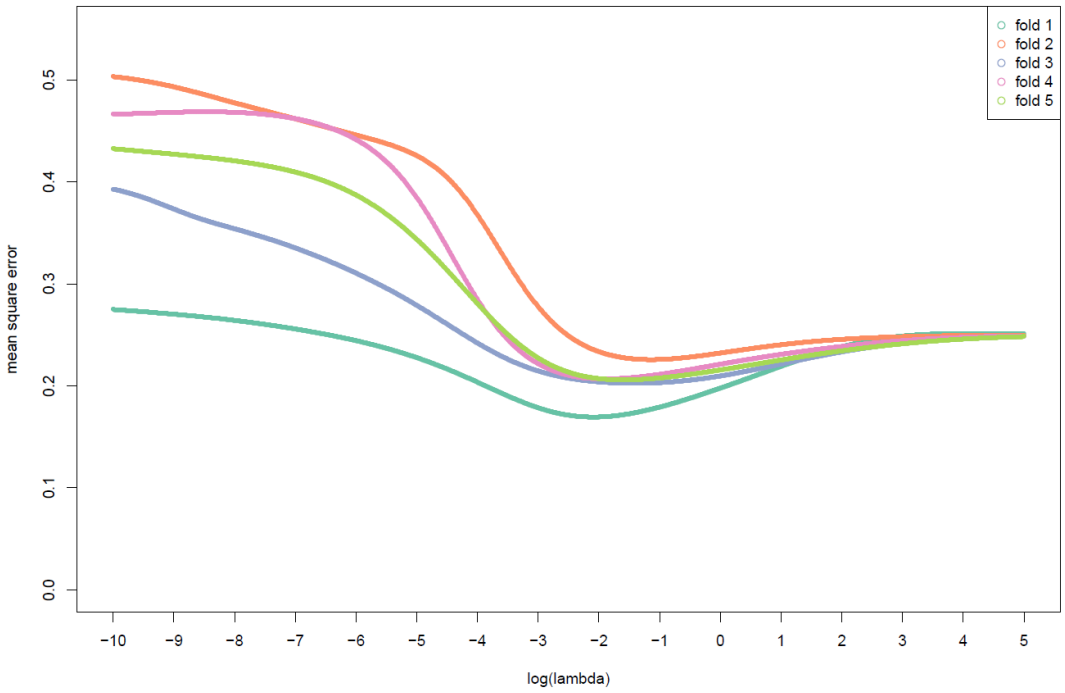
L	Replication, recombination and repair	11.4%	0.99	0.01	yes	enriched
M	Cell wall/membrane/envelope biogenesis	5.4%	0.81	0.19	no	
N	Cell motility	0.0%	0.81	0.19	no	
O	Post-translational modification, protein turnover, and chaperones	0.0%	0.04	0.96	yes	depleted
P	Inorganic ion transport and metabolism	0.6%	0.01	0.99	yes	depleted
Q	Secondary metabolites biosynthesis, transport, and catabolism	0.0%	0.24	0.75	no	
S	Function unknown	28.7%	0.91	0.09	no	
T	Signal transduction mechanisms	3.0%	0.92	0.08	no	
U	Intracellular trafficking, secretion, and vesicular transport	0.0%	0.38	0.62	no	
V	Defense mechanisms	1.8%	0.33	0.67	no	
W	Extracellular structures	0.0%	0.97	0.03	yes	enriched

5) Supplementary Table S4. Detailed phage typing results showing the percentage of resistant, weakly susceptible and strongly susceptible bacterial strains per phage preparation.

phage preparation	resistant	weakly susceptable	strongly susceptable

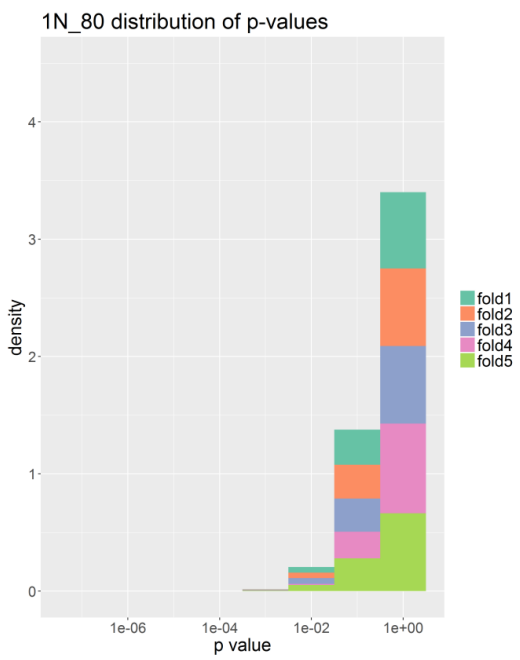
1N/80	45.9	22.2	31.9
676/F	42.5	6.8	50.7
676/T	30.9	1.0	68.1
676/Z	55.6	3.9	40.6
A3/R	77.8	3.4	18.8
A5/L	51.2	1.4	47.3
A5/80	40.1	4.8	55.1
P4/6409	58.9	3.4	37.7
phi200/6409	45.9	10.1	44.0
MS-1	58.9	7.2	33.8
OP_MS-1	54.6	6.8	38.6
OP_MS-1_TOP	50.7	9.7	39.6

6) Supplementary Figure S1.
Cumulative mean square error of the inner cross validation vs strength of ridge penalty per outer fold for the model of phage phi200/6409. It can be seen that the minimum error coincides at similar lambda values for the five folds. Other phage models behaved in comparable fashion.

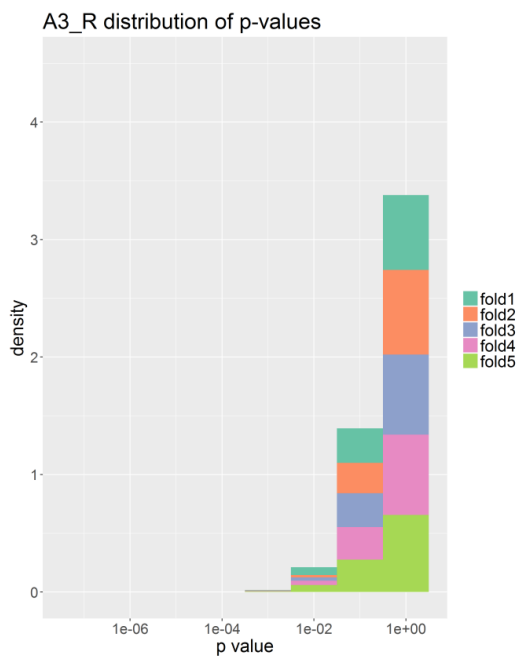


7) Supplementary Figure S2: P-value distributions of gene enrichment analysis on phage preparations a) 1N_80, b) A3_R and c) cocktail MS-1. It can be seen that there is no tail of low p-values as observed for the other phages (compare Figure 3) and the distributions resemble more closely that of the permuted data for the other phages.

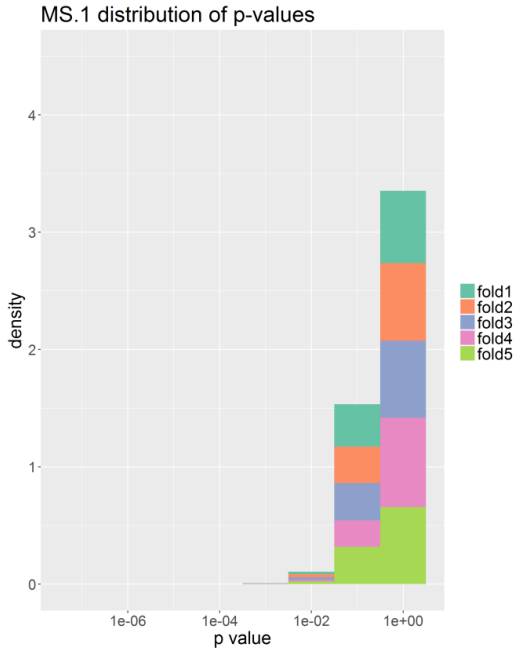
a)


1N_80 distribution of p-values
density
p value
1e-06
1e-04
1e-02
1e+00
0
1
2
3
4
fold1
fold2
fold3
fold4
fold5

b)


A3_R distribution of p-values
density
p value
1e-06
1e-04
1e-02
1e+00
0
1
2
3
4
fold1
fold2
fold3
fold4
fold5

c)


MS.1 distribution of p-values
density
p value
1e-06
1e-04
1e-02
1e+00
0
1
2
3
4
fold1
fold2
fold3
fold4
fold5

Bibliography

[1] Forest Rohwer and Rob Edwards. The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of bacteriology*, 184(16):4529–4535, 2002. 3, 5

[2] Martha Rj Clokie, Andrew D Millard, Andrey V Letarov, and Shaun Heaphy. Phages in nature. *Bacteriophage*, 1(1):31–45, jan 2011. 4

[3] Hans W. Ackermann. Classification of Bacteriophages. In Richard Calendar and Stephen T. Abedon, editors, *The Bacteriophages*, chapter 2, pages 8–16. Oxford University Press, 2 edition, 2006. 4, 5

[4] Richard Calendar and ST Abedon. *The bacteriophages.* Oxford University Press, 2 edition, 2006. 4

[5] Graham F Hatfull. Bacteriophage Genomics. *Current opinion in microbiology*, 11(5):447–453, 2008. 5

[6] George P. C. Salmond and Peter C. Fineran. A century of the phage: past, present and future. *Nature reviews. Microbiology*, 13(12):777–86, 2015. 5

[7] Graham F Hatfull and Roger W Hendrix. Bacteriophages and their Genomes. *Current opinion in virology*, 1(4):298–303, 2011. 5

[8] Bonnie L. Hurwitz, Jana M. U'Ren, and Ken Youens-Clark. Computational prospecting the great viral unknown. *FEMS Microbiology Letters*, 363(10), may 2016. 5

[9] Jean Paul Pirnay, Daniel De Vos, Gilbert Verbeken, Maia Merabishvili, Nina Chanishvili, Mario Vaneechoutte, Martin Zizi, Geert Laire, Rob Lavigne, Isabelle Huys, Guy Van Den Mooter, Angus Buckling, Laurent Debarbieux, Flavie Pouillot, Joana Azeredo, Elisabeth Kutter, Alain Dublanchet, Andrzej Górski, and Revaz Adamia. The phage therapy paradigm: Prêt-à-porter or sur-mesure? *Pharmaceutical Research*, 28(4):934–937, 2011. 7, 91, 92

[10] Simon J. Labrie, Julie E. Samson, and Sylvain Moineau. Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*, 8(5):317–327, mar 2010. 7, 8

[11] K Nordström and A Forsgren. Effect of protein A on adsorption of bacteriophages to Staphylococcus aureus. *Journal of virology*, 14(2):198–202, aug 1974. 7

[12] Eric S Miller, Elizabeth Kutter, Gisela Mosig, Takashi Kunisawa, Wolfgang Rüger, Fumio Arisaka, and Wolfgang Ru. Bacteriophage T4 Genome. *Microbiology and Molecular Biology Reviews*, 67(1):86–156, 2003.

[13] Paul Hyman and Stephen T Abedon. Bacteriophage host range and bacterial resistance. *Advances in applied microbiology*, 70:217–48, 2010.

[14] David Schwarzer, Falk F R Buettner, Christopher Browning, Sergey Nazarov, Wolfgang Rabsch, Andrea Bethe, Astrid Oberbeck, Valorie D Bowman, Katharina Stummeyer, Martina Mühlenhoff, Petr G Leiman, and Rita Gerardy-Schahn. A multivalent adsorption apparatus explains the broad host range of phage phi92: a comprehensive genomic and structural analysis. *Journal of virology*, 86(19):10384–98, oct 2012. 7

[15] Guoqing Xia and Christiane Wolz. Phages of Staphylococcus aureus and their impact on host evolution. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 21:593–601, jan 2014. 7, 13

[16] Nina Molin Høyland-Kroghsbo, Rasmus Baadsgaard Maerkedahl, and Sine Lo Svenningsen. A quorum-sensing-induced bacteriophage defense mechanism. *mBio*, 4(1):e00362–12, feb 2013. 8, 91

[17] Samuel Fischer, Sophie Kittler, Günter Klein, and Gerhard Glünder. Impact of a Single Phage and a Phage Cocktail Application in Broilers on Reduction of Campylobacter jejuni and Development of Resistance. *PLoS ONE*, 8(10), 2013. 8

[18] Hiroya Kunisaki and Yasunori Tanji. Intercrossing of phage genomes in a phage cocktail and stable coexistence with Escherichia coli O157:H7 in anaerobic continuous culture. *Applied Microbiology and Biotechnology*, 85(5):1533–1540, feb 2010. 8

[19] D R Harper, J Anderson, and M C Enright. Phage therapy: delivering on the promise. *Therapeutic delivery*, 2(7):935–47, jul 2011. 8

[20] IITD. Phage Therapy Unit of the Medical Centre of the Institute of Immunology and Experimental Therapy PAS. Available at `https://www.iitd.pan.wroc.pl/en/OTF`. 10

[21] Maya Merabishvili, Jean-Paul Pirnay, Gilbert Verbeken, Nina Chanishvili, Marina Tediashvili, Nino Lashkhi, Thea Glonti, Victor Krylov, Jan Mast, Luc Van Parys, Rob Lavigne, Guido Volckaert, Wesley Mattheus, Gunther Verween, Peter De Corte, Thomas Rose, Serge Jennes, Martin Zizi, Daniel De Vos, and Mario Vaneechoutte. Quality-controlled small-scale production of a well-defined bacteriophage cocktail for use in human clinical trials. *PloS one*, 4(3):e4944, jan 2009. 10

[22] Callum J. Cooper, Mohammadali Khan Mirzaei, and Anders S. Nilsson. Adapting drug approval pathways for bacteriophage-based therapeutics. *Frontiers in Microbiology*, 7(AUG):1–15, 2016. 10

[23] Mark Zipkin. FDA's phage philosophy. Available at `https://www.biocentury.com/bc-innovations/strategy/2017-08-03/fda-forges-paths-phage-therapies`, 2017. 10

[24] Timothy Foster. Staphylococcus. In Samuel Baron, editor, *Medical Microbiology*, chapter 12. University of Texas Medical Branch at Galveston, Galveston, 4th edition, 1996. 13

[25] Alex van Belkum, Damian C Melles, Jan Nouwen, Willem B van Leeuwen, Willem van Wamel, Margreet C Vos, Heiman F L Wertheim, and Henri A Verbrugh. Co-evolutionary aspects of human colonisation and infection by Staphylococcus aureus. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 9(1):32–47, jan 2009. 13

[26] Katherine O'Riordan and Jean C Lee. Staphylococcus aureus capsular polysaccharides. *Clinical microbiology reviews*, 17(1):218–34, jan 2004. 13

[27] Rosanna Capparelli, Marianna Parlato, Giorgia Borriello, Paola Salvatore, and Domenico Iannelli. Experimental phage therapy against Staphylococcus aureus in mice. *Antimicrobial agents and chemotherapy*, 51(8):2765–73, aug 2007.

[28] Marie Deghorain and Laurence Van Melderen. The Staphylococci Phages Family: An Overview. *Viruses*, 4(12):3316–3335, nov 2012. 13, 14

[29] Ruud H Deurenberg and Ellen E Stobberingh. The evolution of Staphylococcus aureus. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 8(6):747–63, dec 2008. 13

[30] Yves Gillet, Bertrand Issartel, Philippe Vanhems, Jean-Christophe Fournet, Gerard Lina, Michèle Bes, François Vandenesch, Yves Piémont, Nicole Brousse, Daniel Floret, and Jerome Etienne. Association between Staphylococcus aureus strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. *Lancet*, 359(9308):753–9, mar 2002. 13

[31] H.P.J. Buermans and J.T. den Dunnen. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10):1932–1941, 2014. 15

[32] Christoph Bleidorn. Third generation sequencing: Technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, 14(1):1–8, 2016. 15, 17

[33] Erwin L. van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, aug 2014. 15, 20

[34] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 2012:251364, 2012. 15

[35] Jared M Churko, Gary L Mantalas, Michael P Snyder, and Joseph C Wu. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circulation research*, 112(12):1613–23, jun 2013. 16

[36] James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 2016. 17

[37] Thomas Hackl, Rainer Hedrich, J??rg Schultz, and Frank F??rster. Proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, 30(21):3004–3011, nov 2014. 17

[38] Simon Andrews. FastQC - A quality control tool for high throughput sequence data. Available at `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. 19

[39] Robert Schmieder and Robert Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, 27(6):863–4, mar 2011. 19

[40] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–9, may 2008. 19

[41] Sergey Nurk, Anton Bankevich, Dmitry Antipov, Alexey Gurevich, Anton Korobeynikov, Alla Lapidus, Andrey Prjibelsky, Alexey Pyshkin, Alexander Sirotkin, Yakov Sirotkin, Ramunas Stepanauskas, Jeffrey McLean, Roger Lasken, Scott R. Clingenpeel, Tanja Woyke, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. Assembling genomes and mini-metagenomes from highly chimeric reads. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7821 LNBI, pages 158–170. Springer, Berlin, Heidelberg, 2013. 19

[42] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, jul 2009. 19

[43] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10, oct 1990. 19

[44] Ea Zankari, Henrik Hasman, Salvatore Cosentino, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M Aarestrup, and Mette Voldby Larsen. Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy*, 67(11):2640–4, nov 2012. 19

[45] Katrine Grimstrup Joensen, Flemming Scheutz, Ole Lund, Henrik Hasman, Rolf S Kaas, Eva M Nielsen, and Frank M Aarestrup. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *Journal of clinical microbiology*, 52(5):1501–10, may 2014. 19

[46] Henrik Hasman, Dhany Saputra, Thomas Sicheritz-Ponten, Ole Lund, Christina Aaby Svendsen, Niels Frimodt-Moller, and Frank M Aarestrup. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *Journal of Clinical Microbiology*, 52(1):139–146, jan 2014. 19

[47] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, dec 2012. 20

[48] J. Besemer. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, 29(12):2607–2618, jun 2001. 20

[49] Ramy K Aziz, Daniela Bartels, Aaron a Best, Matthew DeJongh, Terrence Disz, Robert a Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, Folker Meyer, Gary J Olsen, Robert Olson, Andrei L Osterman, Ross a Overbeek, Leslie K McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko. The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, 9:75, jan 2008. 20

[50] National Research Council. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press, Washington, DC, 2007. 20

[51] J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society*, 135(3):370–384, 1972. 21, 24

[52] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Model Assessment and Selection. In *The Elements of Statistical Learning*, chapter 7. Springer, 2001. 22

[53] Kelly H Zou, A James O'Malley, and Laura Mauri. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–7, feb 2007. 23

[54] Gary C. McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, jul 2009. 24

[55] Isabelle Guyon and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, 3(3):1157–1182, 2003. 26

[56] A Sulakvelidze, Z Alavidze, and J G Morris. Bacteriophage therapy. *Antimicrobial agents and chemotherapy*, 45(3):649–59, mar 2001. 29

[57] Malgorzata Lobocka, Monika S. Hejnowicz, U. Gkagała, Beata Weber-Dabrowska, G. Wkegrzyn, and Michal Dadlez. The first step to bacteriophage therapy – how to choose the correct phage. In Jan Borysowski, Ryszard Miedzybrodzki, and Andrzej Górski, editors, *Phage Therapy: Current Research and Applications*. Norfolk: Caister Academic Press, 2014. 51

[58] Bas E. Dutilh, Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, Daan R. Speth, Victor Seguritan, Ramy K. Aziz, Ben Felts, Elizabeth a. Dinsdale, John L. Mokili, and Robert a. Edwards. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications*, 5:1–11, jul 2014. 90

[59] FDA. Transcript of FDA workshop 'Bacteriophage Therapy: Scientific and Regulatory Issues'. Available at `https://www.fda.gov/BiologicsBloodVaccines/NewsEvents/WorkshopsMeetingsConferences/ucm544294.htm`, 2017. 92

[60] Elizabeth Kutter, Daniel De Vos, Guram Gvasalia, Zemphira Alavidze, Lasha Gogokhia, Sarah Kuhl, and Stephen T Abedon. Phage therapy in clinical practice: treatment of human infections. *Current pharmaceutical biotechnology*, 11(1):69–86, 2010. 92